DOCUMENT RESUME

ED 406 202 SE 059 917

AUTHOR	Johnson, Eugene G.; And Others
TITLE	Technical Report of the NAEP 1992 Trial State Assessment Program in Mathematics.
INSTITUTION	Educational Testing Service, Princeton, NJ. Center for the Assessment of Educational Progress.; National Assessment of Educational Progress, Princeton, NJ.
SPONS AGENCY	National Center for Education Statistics (ED), Washington, DC.
REPORT NO	ISBN-0-88685-139-4; NAEP-23-ST-05
PUB DATE	Apr 93
NOTE	439p.
AVAILABLE FROM	Education Information Branch, Office of Educational Research and Improvement, U.S. Department of Education, 555 New Jersey Avenue, N.W., Washington, DC 20208-5641.
PUB TYPE	Statistical Data (110) -- Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC18 Plus Postage.
DESCRIPTORS	*Academic Achievement; Elementary Education; Evaluation; *Mathematics Achievement; *Research Methodology; Tables (Data)
IDENTIFIERS	*National Assessment of Educational Progress; *Trial State Assessment (NAEP)

ABSTRACT

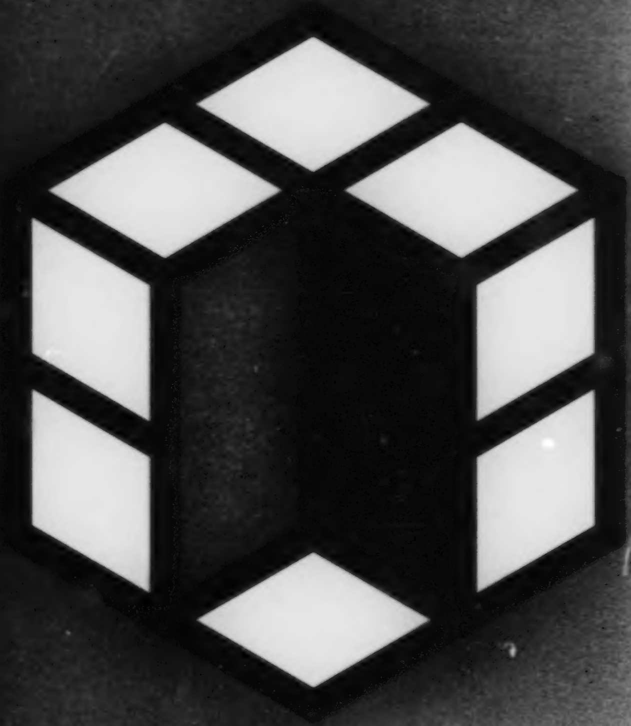
The "Nation's Report Card," the National Assessment of Educational Progress (NAEP), is the only nationally representative and continuing assessment of what America's students know and can do in various subject areas. This report summarizes some of the sophisticated statistical methodology used in the 1992 Trial State Assessment of Mathematics. Chapters include: (1) "Overview: The Design, Implementation, and Analysis of the Trial State Mathematics Assessment Program" (Eugene G. Johnson, Stephen L. Koffler, and John Mazzeo); (2) "Developing the Mathematics Objectives, Cognitive Items, Background Questions, and Assessment Instruments" (Stephen L. Koffler); (3) "Sample Design and Selection" (Leyla K. Mohadjer, Keith F. Rust, Valerija Smith, and Jacqueline Severynse); (4) "State and School Cooperation and Field Administration" (Nancy Caldwell); (5) "Processing Assessment Materials" (Dianne Smrdel, Linda Reynolds, and Brad Thayer); (6) "Creation of the Database and Evaluation of the Quality Control of Data Entry" (John J. Ferris and David S. Freund); (7) "Weighting Procedures and Variance Estimation" (Adam Chu and Keith F. Rust); (8) "Theoretical Background and Philosophy of NAEP Scaling Procedures" (Eugene G. Johnson, Robert J. Mislevy, and Neal Thomas); (9) "Data Analysis and Scaling for the 1992 Trial State Assessment in Mathematics" (John Mazzeo, Huahua Chang, Edward Kulick, Y. Fai Fong, and Angela Grima); and (10) "Conventions Used in Reporting the Results of the 1992 Trial State Assessment in Mathematics" (John Mazzeo). Contains extensive appendixes and 68 references. (JRH)

SE

ED 406 202

NATIONAL CENTER FOR EDUCATION STATISTICS

Technical Report of the NAEP 1992 Trial State Assessment Program in Mathematics



U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

SE053917

THE NATION'S
REPORT
CARD
naep

[illegible]

BEST COPY AVAILABLE

What is The Nation's Report Card?

THE NATION'S REPORT CARD, the National Assessment of Educational Progress (NAEP), is the only nationally representative and continuing assessment of what America's students know and can do in various subject areas. Since 1969, assessments have been conducted periodically in reading, mathematics, science, writing, history/geography, and other fields. By making objective information on student performance available to policymakers at the national, state, and local levels, NAEP is an integral part of our nation's evaluation of the condition and progress of education. Only information related to academic achievement is collected under this program. NAEP guarantees the privacy of individual students and their families.

NAEP is a congressionally mandated project of the National Center for Education Statistics, the U.S. Department of Education. The Commissioner of Education Statistics is responsible, by law, for carrying out the NAEP project through competitive awards to qualified organizations. NAEP reports directly to the Commissioner, who is also responsible for providing continuing reviews, including validation studies and solicitation of public comment, on NAEP's conduct and usefulness.

In 1988, Congress created the National Assessment Governing Board (NAGB) to formulate policy guidelines for NAEP. The board is responsible for selecting the subject areas to be assessed, which may include adding to those specified by Congress; identifying appropriate achievement goals for each age and grade; developing assessment objectives; developing test specifications; designing the assessment methodology; developing guidelines and standards for data analysis and for reporting and disseminating results; developing standards and procedures for interstate, regional, and national comparisons; improving the form and use of the National Assessment; and ensuring that all items selected for use in the National Assessment are free from racial, cultural, gender, or regional bias.

The National Assessment Governing Board

Mark D. Musick, Chairman
President
Southern Regional Education Board
Atlanta, Georgia

Hon. William T. Randall, Vice Chair
Commissioner of Education
State Department of Education
Denver, Colorado

Parris C. Battle
Education Specialist
Dade County Public Schools
Miami, Florida

Honorable Evan Bayh
Governor of Indiana
Indianapolis, Indiana

Mary R. Blanton
Attorney
Blanton & Blanton
Salisbury, North Carolina

Boyd W. Boehlje
Attorney and School Board Member
Pella, Iowa

Linda R. Bryant
Dean of Students
Florence Reizenstein Middle School
Pittsburgh, Pennsylvania

Naomi K. Cohen
Office of Policy and Management
State of Connecticut
Hartford, Connecticut

Charlotte Crabtree
Professor
University of California
Los Angeles, California

Chester E. Finn, Jr.
Founding Partner and Senior Scholar
The Edison Project
Washington, DC

Michael S. Glode
Wyoming State Board of Education
Saratoga, Wyoming

William Hume
Chairman of the Board
Basic American, Inc.
San Francisco, California

Christine Johnson
Director of K-12 Education
Littleton Public Schools
Littleton, Colorado

John S. Lindley
Principal
Galloway Elementary School
Henderson, Nevada

Honorable Stephen E. Merrill
Governor of New Hampshire
Concord, New Hampshire

Jason Millman
Professor
Cornell University
Ithaca, New York

Honorable Richard P. Mills
Commissioner of Education
State Department of Education
Montpelier, Vermont

Carl J. Moser
Director of Schools
The Lutheran Church — Missouri Synod
St. Louis, Missouri

John A. Murphy
Superintendent of Schools
Charlotte-Mecklenburg Schools
Charlotte, North Carolina

Michael T. Nettles
Professor
University of Michigan
Ann Arbor, Michigan

Honorable Carolyn Pollan
Arkansas House of Representatives
Fort Smith, Arkansas

Thomas Topuzes
Senior Vice President
Valley Independent Bank
El Centro, California

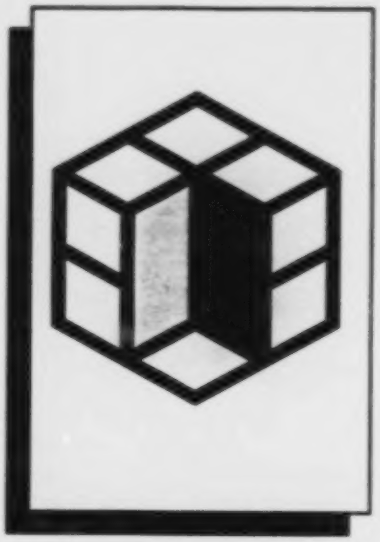
Marilyn Whirry
English Teacher
Mira Costa High School
Manhattan Beach, California

Emerson J. Elliott
Acting Assistant Secretary for Educational Research and Improvement (Ex-Officio)
U.S. Department of Education
Washington, D.C.

Roy Truby
Executive Director, NAGB
Washington, D.C.

NATIONAL CENTER FOR EDUCATION STATISTICS

Technical Report of the NAEP 1992 Trial State Assessment Program in Mathematics



Eugene G. Johnson
John Mazzeo
Debra L. Kline
in collaboration with

Mary Lyn Bourque
Nancy Caldwell
Huahua Chang
Adam Chu
John J. Ferris
Y. Fai Fong
David S. Freund
Angela Grima
Stephen L. Koffler
Edward Kulick
Robert J. Mislevy
Leyla K. Mohadjer
Ina V. S. Mullis
Linda Reynolds
Keith F. Rust
Jacqueline Severynse
Valerija Smith
Dianne Smrdel
Brad Thayer
Neal Thomas

with a Foreword by Gary W. Phillips

Report No. 23-ST05 April 1993

THE NATION'S REPORT CARD

Prepared by Educational Testing Service under contract with the National Center for Education Statistics

Office of Educational Research and Improvement
U.S. Department of Education

U.S. Department of Education
Richard W. Riley
Secretary

Office of Educational Research and Improvement
Emerson J. Elliott
Acting Assistant Secretary

National Center for Education Statistics
Emerson J. Elliott
Commissioner

FOR MORE INFORMATION:

For ordering information on this report, write:

Education Information Branch
Office of Educational Research and Improvement
U.S. Department of Education
555 New Jersey Avenue, NW
Washington, D.C. 20208-5641

or call 1-800-424-1616 (in the Washington, D.C. metropolitan area call 202-219-1651).

Library of Congress, Catalog Card Number: 93-83070

ISBN: 0-88685-139-4

The work upon which this publication is based was performed for the National Center for Education Statistics, Office of Educational Research and Improvement, by Educational Testing Service.

Educational Testing Service is an equal opportunity, affirmative action employer.

Educational Testing Service, ETS, and the ETS logo are registered trademarks of Educational Testing Service.

TECHNICAL REPORT
OF THE NAEP 1992 TRIAL STATE ASSESSMENT PROGRAM
IN MATHEMATICS

TABLE OF CONTENTS

List of Tables and Figures ix

Acknowledgments xiii

Foreword xvii
Gary W. Phillips

Chapter 1 Overview: The Design, Implementation, and Analysis of the Trial State Mathematics Assessment Program 1
Eugene G. Johnson, Stephen L. Koffler, and John Mazzeo

1.1 Overview 1
1.2 Design of the Trial State Assessment in Mathematics 6
1.3 Development of Mathematics Objectives, Items, and Background Questions 7
1.4 Assessment Instruments 8
1.5 The Sampling Design 9
1.6 Field Administration 9
1.7 Materials Processing and Database Creation 10
1.8 The Trial State Assessment Data 11
1.9 Weighting and Variance Estimation 11
1.10 Preliminary Data Analysis 12
1.11 Scaling the Assessment Items 13
1.12 Linking the Trial State Results to the National Results 14
1.13 Reporting the Trial State Assessment Results 15

Chapter 2 Developing the Mathematics Objectives, Cognitive Items, Background Questions, and Assessment Instruments 17
Stephen L. Koffler

2.1 Overview 17
2.2 Context for Planning the 1992 Mathematics Assessment 18
2.3 Assessment Design Principles 18
2.4 Assessment Development Process 19
2.5 Mathematics Framework 20
2.6 Cognitive Item Development 21
2.7 Block Design 24
2.8 Student Assessment Booklets 30

2.9 Questionnaires 31
  2.9.1 Student Questionnaires 32
  2.9.2 Teacher, School, and Excluded Student Questionnaires 33

Chapter 3 Sample Design and Selection 35
Leyla K. Mohadjer, Keith F. Rust, Valerija Smith, and Jacqueline Severynse

3.1 Introduction and Overview 35
3.2 Sample Selection for the 1991 Field Test 37
  3.2.1 Primary Sampling Units 37
  3.2.2 Selection of Schools and Students 38
  3.2.3 Assignment to Sessions for Different Subjects 38
3.3 Sampling Frame for the 1992 Assessment 39
  3.3.1 Choice of School Sampling Frame 39
  3.3.2 Missing Minority and Urbanization Data 40
  3.3.3 In-scope Schools 40
3.4 Within-state Stratification 40
  3.4.1 Stratification Variables 40
  3.4.2 Urbanization Classification 43
  3.4.3 Minority Classification 43
  3.4.4 Median Household Income 66
  3.4.5 Schools With Fewer Than 20 Students 66
3.5 School Sample Selection for the 1992 Trial State Assessment 67
  3.5.1 Control of Overlap of School Samples for National Educational Studies 67
  3.5.2 Selection of Schools in Small States 70
  3.5.3 States with Geographic Clustering of Small Schools 70
  3.5.4 States with Stratification of Small Schools 70
  3.5.5 Overlap of School Samples 70
  3.5.6 New School Selection 72
  3.5.7 Assigning Subject Session Types at Grade 4 74
  3.5.8 Designating Monitor Status 74
  3.5.9 Substitutes 77
3.6 Student Sample Selection 78

Chapter 4 State and School Cooperation and Field Administration 87
Nancy Caldwell

4.1 Overview 87
4.2 The Field Test 87
  4.2.1 Conduct of the Field Test 87
  4.2.2 Results of the Field Test 89
4.3 The 1992 Trial State Assessment 89
  4.3.1 Overview of Responsibilities 89
  4.3.2 Schedule of Data Collection Activities 92
  4.3.3 Preparations for the Trial State Assessment 93
  4.3.4 Monitoring of Assessment Activities 96
  4.3.5 School and Student Participation 96
  4.3.6 Results of the Observations 98

Chapter 5 Processing Assessment Materials 101
Dianne Smrdel, Linda Reynolds, and Brad Thayer

5.1 Overview 101
5.2 Process Control System 102
5.3 Workflow Management System 102
5.4 Process Flow of NAEP Materials and Database Creation 103
5.5 Materials Distribution 103
5.6 Processing Assessment Materials 107
5.7 Professional Scoring 110
  5.7.1 Description of Scoring 110
  5.7.2 Training 111
  5.7.3 Trend Scoring of 1990 Items 113
  5.7.4 Reliability of Scoring 113
5.8 Data Transcription Systems 115
  5.8.1 Data Entry 115
  5.8.2 Scanning 115
5.9 Data Validation 116
5.10 Editing 117
5.11 Questionnaires 119
5.12 Merging of Student Data 119
5.13 Storage of Documents 120

Chapter 6 Creation of the Database and Evaluation of the Quality Control of Data Entry 121
John J. Ferris and David S. Freund

6.1 Overview 121
6.2 Merging Files into the Trial State Assessment Database 121
6.3 Creating the Master Catalog 122
6.4 Quality Control Evaluation 123
  6.4.1 Student Data 123
  6.4.2 Teacher Questionnaires 125
  6.4.3 School Questionnaires 126
  6.4.4 Excluded Student Questionnaires 126

Chapter 7 Weighting Procedures and Variance Estimation 129
Adam Chu and Keith F. Rust

7.1 Introduction 129
7.2 Calculation of Base Weights 130
  7.2.1 Calculation of School/hit Base Weights 130
  7.2.2 Weighting New Schools 131
  7.2.3 Treatment of Substitute and Double-session Substitute Schools 132
  7.2.4 Calculation of Student Base Weights 132
7.3 Adjustments for Nonresponse 133
  7.3.1 Defining Initial School-level Nonresponse Adjustment Classes 133
  7.3.2 Constructing the Final Nonresponse Adjustment Classes 134
  7.3.3 School/hit Adjustment Factors 135

7.3.4 Student-level Nonresponse Adjustment Classes 136
7.3.5 Student Nonresponse Adjustments 137
7.4 Characteristics of Nonresponding Schools and Students 138
7.4.1 Weighted Distributions of Schools Before and After Nonresponse 141
7.4.2 Characteristics of Nonresponding Schools 141
7.4.3 Weighted Distributions of Students Before and After Student Absenteeism 149
7.4.4 Characteristics of Absent Students 152
7.5 Variation in Weights 159
7.6 Calculation of Replicate Weights 160
7.6.1 Defining Replicate Groups for Variance Estimation 161
7.6.2 School-level Replicate Weights 162
7.6.3 Student-level Replicate Weights 163
7.7 Calculation of School Weights 165

Chapter 8 Theoretical Background and Philosophy of NAEP Scaling Procedures 167
Eugene G. Johnson, Robert J. Mislevy, and Neal Thomas

8.1 Overview 167
8.2 Background 167
8.3 Scaling Methodology 169
8.3.1 The Scaling Models 169
8.3.2 An Overview of Plausible Values Methodology 173
8.3.3 Computing Plausible Values in IRT-based Scales 175
8.4 Analyses 176
8.4.1 Computational Procedures 176
8.4.2 Statistical Tests 178
8.4.3 Biases in Secondary Analyses 178
8.5 Scale Anchoring and Achievement Levels 179

Chapter 9 Data Analysis and Scaling for the 1992 Trial State Assessment in Mathematics 181
John Mazzeo, Huahua Chang, Edward Kulick, Y. Fai Fong, and Angela Grima

9.1 Overview 181
9.2 Description of Items, Assessment Booklets, and Administration Procedures 182
9.3 Item Analyses 186
9.3.1 Conventional Item and Test Analyses 186
9.3.2 Differential Item Functioning (DIF) Analyses 194
9.4 Item Response Theory (IRT) Scaling 198
9.4.1 Item Parameter Estimation 208
9.5 Estimation of State and Subgroup Proficiency Distributions 217
9.6 Linking State and National Scales 225
9.7 Producing a Mathematics Composite Scale 238

Chapter 10 Conventions Used in Reporting the Results of the 1992 Trial State Assessment in Mathematics 241
John Mazzeo

10.1 Overview 242
10.2 Minimum Sample Sizes for Reporting Subgroup Results 243
10.3 Estimates of Standard Errors with Large Mean Squared Errors 244
10.4 Treatment of Missing Data from the Questionnaires 245
10.5 Statistical Rules Used for Producing the State Reports 248
  10.5.1 Comparing Means and Proportions for Mutually Exclusive Groups of Students 249
  10.5.2 Multiple Comparison Procedure 250
  10.5.3 Determining the Highest and Lowest Scoring Groups from a Set of Ranked Groups 250
  10.5.4 Comparing 1992 and 1990 Results in State Report Tables 251
  10.5.5 Comparing Dependent Proportions 252
  10.5.6 Statistical Significance and Estimated Effect Sizes 253
  10.5.7 Descriptions of the Magnitude of Percentage 254
10.6 Comparisons of 1992 and 1990 Eighth-grade Results in the *Mathematics Report Card* and the *Data Compendium* 254

• • •

Appendix A Participants in the Objectives and Item Development Process 261

Appendix B Summary of Participation Rates 267

Appendix C Conditioning Variables and Contrast Codings 291

Appendix D IRT Parameters for Mathematics Items 317

Appendix E Trial State Assessment Reporting Subgroups; Composite and Derived Common Background Variables; Composite and Derived Reporting Variables 335

Appendix F The NAEP Scale Anchoring Process for the 1992 Mathematics Assessment 349
Ina V.S. Mullis and Eugene G. Johnson

Appendix G The NAEP Achievement Level Setting Process for the 1992 Mathematics Assessment 367
Mary Lyn Bourque

Appendix H Reanalysis of the 1990 Trial State Assessment Data 391
John Mazzeo

References Cited in Text 401

LIST OF TABLES AND FIGURES

Table 1-1 Jurisdictions participating in the 1992 Trial State Assessment 2

2-1 NAEP mathematics framework: Percentage distribution of items by grade and ability 20
2-2 NAEP mathematics framework: Percentage distribution of items by grade and content area 21
2-3 Number and percentage of questions for the 1990 and 1992 assessments according to the mathematics framework, grade 4 25
2-4 Number and percentage of questions for the 1990 and 1992 assessments according to the mathematics framework, grade 8 26
2-5 Total number of items in the 1992 Trial State Assessment in Mathematics 27
2-6 Number of items contributed by 1990 blocks 28
2-7 Cognitive and noncognitive block composition, grade 4 29
2-8 Cognitive and noncognitive block composition, grade 8 29
2-9 Booklet content at each grade level 31

3-1 Distribution of fourth-grade schools and enrollment as reported in QED 1990 41
3-2 Distribution of eighth-grade schools and enrollment as reported in QED 1990 42
3-3 Distribution of the selected schools by sampling strata, grade 4 44
3-4 Distribution of the selected schools by sampling strata, grade 8 55
3-5 Distribution of sample sizes by school size, with corresponding overlap between grades 71
3-6 Number of schools selected for both national and state samples, by state 73
3-7 Distribution of new schools coming from "large" and "small" districts, grade 4 75
3-8 Distribution of new schools coming from "large" and "small" districts, grade 8 76
3-9 Substitute school counts, grade 4 79
3-10 Substitute school counts, grade 8 80
3-11 Distribution of the grade 4 mathematics school sample by state 81
3-12 Distribution of the grade 8 mathematics school sample by state 82
3-13 Distribution of the grade 4 mathematics student sample and response rates by state 84
3-14 Distribution of the grade 8 mathematics student sample and response rates by state 85

4-1 School participation, 1992 Trial State Assessment 97
4-2 Student participation, 1992 Trial State Assessment in mathematics 97

5-1 Interreader reliabilities for extended constructed-response items 114

6-1 Number of assessment booklets scanned and selected for quality control evaluation 124
6-2 Inference from the quality control evaluation of student data 125
6-3 Inference from the quality control evaluation of questionnaire data 127

7-1 Unweighted and weighted counts of assessed students by state and grade 139
7-2 Unweighted and weighted counts of excluded students with returned questionnaires by state and grade 140
7-3 Weighted mean values derived from sampled schools, grade 4 142

7-4 Weighted mean values derived from sampled schools, grade 8 143
7-5 Grade 4 school nonresponse adjustment classes with adjustment factors greater than 1.25 144
7-6 Grade 8 school nonresponse adjustment classes with adjustment factors greater than 1.25 147
7-7 Weighted student percentages derived from sampled schools, grade 4 150

Table 7-8 Weighted student percentages derived from sampled schools, grade 8 151
7-9 Grade 4 student nonresponse adjustment classes with adjustment factors greater than 1.25 153
7-10 Grade 8 student nonresponse adjustment classes with adjustment factors greater than 1.25 154

9-1 Mathematics block composition by scale and item type, grade 4 183
9-2 Mathematics block composition by scale and item type, grade 8 184
9-3 Descriptive statistics for each block of items by position within test booklet and overall, grade 4 187
9-4 Descriptive statistics for each block of items by position within test booklet and overall, grade 8 188
9-5 Block-level descriptive statistics for monitored and unmonitored sessions, grade 4 191
9-6 Block-level descriptive statistics for monitored and unmonitored sessions, grade 8 192
9-7 Frequency distributions of DIF statistics for grade 4 items grouped by content or skill area 196
9-8 Frequency distributions of DIF statistics for grade 8 items grouped by content or skill area 197
9-9 Extended constructed-response items 210
9-10 Items receiving special treatment 217
9-11 Proportion of proficiency variance accounted for by grade 4 conditioning models 221
9-12 Proportion of proficiency variance accounted for by grade 8 conditioning models 222
9-13 Transformation constants for the grade 4 and grade 8 scales 230
9-14 Weights used for each scale to form grade 4 and grade 8 composites 239

10-1 Weighted percentage of students matched to teacher questionnaire, grade 4 246
10-2 Weighted percentage of students matched to teacher questionnaire, grade 8 247
10-3 Rules for selecting descriptions of percentages 254
10-4 Overall average mathematics proficiency and achievement levels 257
10-5 Average mathematics proficiency by race/ethnicity 259

Figure 4-1 Participating jurisdictions, 1990 and 1992 assessments 90

5-1 Data flow overview, 1992 Trial State Assessment 104
5-2 Materials processing flow, 1992 Trial State Assessment 105
5-3 Packing list, 1992 Trial State Assessment 109
5-4 Sample extended constructed-response item and scoring guide 112

9-1 Stem-and-leaf display of state-by-state differences in average item scores (monitored - unmonitored) 193
9-2 Stem-and-leaf display of average scores for items, by scale, for grade 4 199
9-3 Stem-and-leaf display of average scores for items, by scale, for grade 8 202
9-4 Differences in average item scores (monitored minus unmonitored) plotted against monitored average item scores, grade 4 206
9-5 Differences in average item scores (monitored minus unmonitored) plotted against monitored average item scores, grade 8 207
9-6 Stem-and-leaf display of state-by-state differences in average item score (monitored - unmonitored) for the grade 8 estimation item pool 209
9-7 Plots comparing empirical and model-based estimates of item response functions for binary-scored items exhibiting good model fit 213
9-8 Plot comparing empirical and model-based estimates of item category characteristic curves for a polytomously scored item exhibiting good model fit 214

Figure 9-9 Plots comparing empirical and model-based estimates of item response functions for binary-scored items exhibiting some model misfit 215
9-10 Plot comparing empirical and model-based estimates of item category characteristic curves for a polytomously scored item exhibiting some model fit 216
9-11 Plots comparing empirical and model-based estimates of item response functions for items dropped from scaling due to model misfit 218
9-12 Boxplots of estimated scale correlations, grade 4 223
9-13 Boxplots of estimated scale correlations, grade 8 224
9-14 Plot of mean proficiency versus mean item score, grade 4 226
9-15 Plot of mean proficiency versus mean item score, grade 8 227
9-16 Rootogram comparing proficiency distributions for the Trial State Assessment aggregate sample and the state aggregate comparison sample from the national assessment for the numbers and operations scale 232
9-17 Rootogram comparing proficiency distributions for the Trial State Assessment aggregate sample and the state aggregate comparison sample from the national assessment for the measurement scale 233
9-18 Rootogram comparing proficiency distributions for the Trial State Assessment aggregate sample and the state aggregate comparison sample from the national assessment for the geometry scale 234
9-19 Rootogram comparing proficiency distributions for the Trial State Assessment aggregate sample and the state aggregate comparison sample from the national assessment for the data analysis, statistics, and probability scale 235
9-20 Rootogram comparing proficiency distributions for the Trial State Assessment aggregate sample and the state aggregate comparison sample from the national assessment for the algebra and functions scale 236
9-21 Rootogram comparing proficiency distributions for the Trial State Assessment aggregate sample and the state aggregate comparison sample from the national assessment for the estimation scale 237
9-22 Rootogram comparing proficiency distributions for the Trial State Assessment aggregate sample and the state aggregate comparison sample from the national assessment for the composite scale 240

ACKNOWLEDGMENTS

The design, development, analysis, and reporting of the Trial State Assessment Program was truly a collaborative effort among staff from State Education Agencies, the National Center for Education Statistics (NCES), Educational Testing Service (ETS), Westat, and National Computer Systems (NCS). The program benefitted from the contributions of hundreds of individuals at the state and local levels—Governors, Chief State School Officers, State and District Test Directors, State Coordinators, and district administrators—who tirelessly provided their wisdom, experience, and hard work. Finally, and most importantly, NAEP is grateful to the students and school staff who participated in the Trial State Assessment.

This report documents the design and data analysis procedures behind the 1992 Trial State Assessment in mathematics. It also provides insight into the rationale behind the technical decisions made about the program. The development of this Technical Report and, especially of the Trial State Assessment Program, is the culmination of effort by many individuals who contributed their considerable knowledge, experience, and creativity to the 1990 and 1992 Trial State Assessment Programs.

The 1990 and 1992 Trial State Assessments were funded through the National Center of Education Statistics in the Office of Educational Research and Improvement of the U.S. Department of Education. Emerson Elliott, NCES Commissioner, provided consistent support and guidance. The staff—particularly Gary Phillips, Eugene Owen, Stephen Gorman, Susan Ahmed, Andrew Kolstad, Maureen Treacy, and Sheida White—worked closely and collegially with ETS, Westat, and NCS staff and played a crucial role in all aspects of the program.

The members of the National Assessment Governing Board (NAGB) and NAGB staff provided advice and guidance throughout, and their contractor, American College Testing, worked with various panels in setting the achievement levels, and carried out a variety of analyses related to the levels.

The Chief State School Officers managed the National Assessment Planning Project that resulted in the mathematics framework and objectives for the assessment.

NAEP owes a great deal to the numerous panelists and consultants who worked so diligently on developing the assessment and providing a frame for interpreting the results, including those who helped create the objectives, develop the assessment instruments, set the achievement levels, and provide the anchoring descriptions.

Under the NAEP contract to ETS, Archie Lapointe served as the executive director and Ina Mullis as the project director. John Barone directed the data analysis activities; Jules Goodison, the operational aspects; Chancey Jones and Jeffrey Haberstroh, test development; Kent Ashworth, information services; Eugene Johnson, measurement and research; and John Olson, technical assistance and state services.

ETS and NAEP management have been very supportive of NAEP's technical work. Special thanks go to Gregory Anrig and Nancy Cole as well as to Henry Braun and Charles Davis of ETS research management, and to Archie Lapointe, Ina Mullis, Jules Goodison, and David Hobson of NAEP management.

The guidance of the NAEP Design and Analysis Committee on the technical aspects of NAEP has been outstanding. The members are Sylvia Johnson (chair), Albert Beaton, Jeri Benson, John Carroll, Clifford Clogg, William Cooley, Jeremy Finn, Bert Green, Huynh Huynh, Bengt Muthén, Anthony Nitko, Ingram Olkin, Tej Pandey, and Juliet Shaffer.

The design and data analysis of the 1992 Trial State Assessment Program was primarily the responsibility of the NAEP research and data analysis staff, with significant contributions from the NAEP management, Westat, and NCS staffs. Statistical and psychometric activities were led by John Mazzeo, with consultation from Eugene Johnson and with assistance from Spencer Swinton and Huahua Chang. Angela Grima managed the DIF analysis with assistance from Ira Sample and James Rosso. Major contributions were made by Nancy Allen, James Carlson, John Donoghue, Frank Jenkins, Jo-lin Liang, Eiji Muraki, and Neal Thomas. Robert Mislevy and Ming-mei Wang provided valuable statistical and psychometric advice.

The division of Data Analysis and Technical Research, under the outstanding leadership of John Barone, was responsible for developing the operating systems and carrying out the data analyses. Alfred Rogers and David Freund deserve special recognition for their leadership in developing and maintaining the large and complex NAEP data management systems. Alfred Rogers also deserves special mention for his role in the development of production versions of key analysis and scaling systems. Special thanks also go to David Freund, Bruce Kaplan, Edward Kulick, and John J. Ferris for their continuing roles as leaders and developers of innovative software solutions to NAEP data analysis challenges. The individual state level reports were superbly created through the efforts of the computer-generated reporting team: Laura Jerry, Robert Patrick, Jennifer Nelson, Philip Leung, Bruce Kaplan, and John J. Ferris. Other members of this division who made substantial contributions of their talent, and important contributions to NAEP data analyses, were Drew Bowker, Yim Fai Fong, Steven Isham, Laura Jenkins, Michael Narcowich, Craig Pizzuti, Ira Sample, and Minhwei Wang.

The staff of Westat, Inc. contributed their exceptional talents in all areas of sample design and data collection. Particular recognition is due to Renee Slobasky and Nancy Caldwell for supervising the field operations and to Keith Rust for

developing and supervising the sampling design. Debra Vivari, Dianne Walsh, Leyla Mohadjer, Adam Chu, Valerija Smith, and Jacqueline Severynse undertook major roles in these activities also.

Critical to the program was the contribution of National Computer Systems, Inc., which has been responsible for the printing, distribution, and processing of the assessment materials. The leadership roles of John O'Neill and Judith Moyer are especially acknowledged. Thanks go also to Linda Reynolds, Bradley Thayer, and Dianne Smrdel.

Judith Alfort, Margaret Biriki, Donna Lembeck, Marciline Yates, and Mary Varone are acknowledged for their patience and diligence in typing and proofing the many revisions of this report.

Kent Ashworth was responsible for coordinating the cover design and final printing of this report.

Special thanks go to Debra Kline for organizing, scheduling, editing, motivating, and ensuring the cohesiveness and correctness of the final report.

Special thanks are also due to many individuals for their invaluable assistance in reviewing the reports, especially the editors who improved the text and the data analysts who checked the accuracy of the data.

FOREWORD

This technical report summarizes some of the most sophisticated statistical methodology used in any survey or testing program in the United States. In its 23-year history, the National Assessment of Educational Progress has employed such state-of-the-art techniques as matrix sampling and item response theory models. Today it is the only survey using the advanced plausible values methodology, which uses a multiple imputation procedure in a psychometric context.

The 1992 Trial State Assessment of mathematics followed the same basic design as that used for the 1990 Trial State Assessment. Properties of the 1992 assessment common to the 1990 assessment include: 1) continuing the use of focused-BIB spiraling, item response theory models, and plausible values; 2) keeping the national and Trial State Assessment samples unduplicated; 3) doing separate stratifications and conditioning in each of the state samples; 4) making each state sample have power similar to the regional samples from the national assessment (this is how the sample sizes for the states were determined); 5) equating the aggregate of the state samples to the national scale (and doing this via a national subsample that also was representative of the aggregate of the states); 6) limiting the state samples to public schools; and 7) using power rules to determine which subgroup comparisons were supported by sufficient sample sizes (this became the "rule of 62").

There were several changes in the 1992 effort that should be noted. The most obvious change was the inclusion of an assessment of fourth-grade public-school students in addition to the assessment of eighth-grade students (the only grade assessed in 1990). More items were added to the assessment (many of which were of the non-multiple choice variety) resulting in a change from a 7-block focused-BIB design in 1990 to a 13-block focused-BIB design in 1992. In addition, special booklets of items were administered to measure students' estimation abilities. Another major change was that the National Assessment Governing Board established new within-grade Basic, Proficient, and Advanced achievement levels on the NAEP scale. These were improvements over the 1990 effort, and, in fact, represent the initial and primary way of reporting the 1992 results. Finally, there were some improvements in the conditioning process, which allowed more precise estimation of the correlation between content area scales. These changes necessitated a rescaling of the 1990 data (so that it is on the same scale as 1992) and a reanalysis of the 1990 results. For this reason, the reader of the 1992 reports may see some differences (generally slight) in the old reporting of the 1990 results compared to the new reporting of the 1990 results.

For all the technical people working on the 1992 Trial State Assessment, the NAEP project has tested the limits of statistical theory and provided many opportunities to advance the state of the art.

The NAEP project is not only characterized by its elegant statistical procedures, but it is also noted for the dedicated professionalism of its staff. In hundreds of hours of technical advisory committees, I have not seen a single instance in which truth, honesty, and reason were compromised. It is the stubborn insistence that surveys are scientific activities and the relentless quest for improved methodology that have made NAEP credible for more than 23 years.

Gary W. Phillips
Associate Commissioner
National Center for Education Statistics

Chapter 1

OVERVIEW:

THE DESIGN, IMPLEMENTATION, AND ANALYSIS OF THE 1992 TRIAL STATE ASSESSMENT PROGRAM IN MATHEMATICS

Eugene G. Johnson, Stephen L. Koffler, and John Mazzeo

Educational Testing Service

> *The National Assessment shall develop a trial mathematics assessment survey instrument for the 8th grade and shall conduct a demonstration of the instrument in 1990 in States which wish to participate, with the purpose of determining whether such an assessment yields valid, reliable State representative data. (Section 406 (i)(2)(C)(i) of the General Education Provisions Act, as amended by Pub. L. 100-297 (20 US.C. 1221e-1(i)(2)(C)(i)))*
>
> *The National Assessment shall conduct a trial mathematics assessment for the fourth and eighth grades in 1992 and, pursuant to subparagraph (6)(D), shall develop a trial reading assessment to be administered in 1992 for the fourth grade in States which wish to participate, with the purpose of determining whether such an assessment yields valid, reliable State representative data. (Section 406 (i)(2)(C)(i) of the General Education Provisions Act, as amended by Pub. L. 100-297 (20 US.C. 1221e-1(i)(2)(C)(ii)))*

1.1 OVERVIEW

In April 1988, Congress reauthorized the National Assessment of Educational Progress (NAEP) and added a new dimension to the program—voluntary state-by-state assessments on a trial basis in 1990 and 1992, in addition to continuing the national assessments that NAEP had conducted since its inception. In this report, we will refer to the voluntary state-by-state assessment program as the Trial State Assessment Program. These assessments, which are designed to provide state representative data, are distinct from the assessment designed to provide nationally representative data, referred to in this report as the national assessment. (This terminology is also used in all other reports of the 1990 and 1992 assessments.) It should be noted that the word trial in Trial State Assessment refers to the Congressionally mandated trial to determine whether such assessments can yield valid, reliable state representative data. All instruments and procedures used in the 1990 and 1992 Trial State and national assessments were previously piloted in field tests conducted in the year prior to the assessment.

The 1990 Trial State Assessment Program collected information on the mathematics knowledge, skills, understanding, and perceptions of a representative sample of eighth-grade students in public schools in 37 states, the District of Columbia, and two territories. The second phase of the Trial State Assessment Program, conducted in 1992, collected information on the mathematics knowledge, skills, understanding, and perceptions of a representative sample of fourth- and eighth-grade students and the reading knowledge, skills, understanding, and perceptions of a representative sample of fourth-grade students in public schools in 41 states, the District of Columbia, and two territories.[1]

Table 1-1 lists the jurisdictions that participated in the 1992 Trial State Assessment Program. About 110,000 students at each grade level participated in the mathematics assessments in those jurisdictions. The students who were assessed in mathematics were administered one of 26 mathematics assessment booklets also used in NAEP's 1992 national mathematics assessment. In addition, all students participating in the Trial State Assessment in mathematics were given a special set of questions measuring estimation skills that was also administered as part of the national program. The estimation block was administered using a special audiotape to pace the students through the items.

Table 1-1
Jurisdictions Participating in the
1992 Trial State Assessment Program

Jurisdictions			
Alabama	Hawaii	_Mississippi_*	Pennsylvania
Arizona	Idaho	_Missouri_*	Rhode Island
Arkansas	Indiana	Nebraska	_South Carolina_*
California	Iowa	New Hampshire	_Tennessee_*
Colorado	Kentucky	New Jersey	Texas
Connecticut	Louisiana	New Mexico	_Utah_*
Delaware	_Maine_*	New York	Virginia
District of Columbia	Maryland	North Carolina	Virgin Islands**
Florida	_Massachusetts_*	North Dakota	West Virginia
Georgia	Michigan	Ohio	Wisconsin
Guam	Minnesota	Oklahoma	Wyoming

* These states did not participate in the 1990 Trial State Assessment Program. Illinois, Montana, and Oregon participated in the 1990 program but did not participate in the 1992 program.

** The Virgin Islands participated in the 1992 Trial State Assessment Program. However, in accordance with the legislation providing for participants to review and give permission for release of their results, the Virgin Islands chose not to publish their results at grade 4.

---

[1] This report outlines the technical details of the 1992 Trial State Assessment in mathematics. A separate report on the technical details of the 1992 Trial State Assessment in reading will be published at the same time as the results from the reading assessment.

The mathematics framework and objectives established to guide both the Trial State Assessment and national assessment were developed for NAEP through a consensus project of the Council of Chief State School Officers, funded by the National Center for Education Statistics and the National Science Foundation. The framework and objectives were also used for the 1990 and 1992 national mathematics assessments. In addition, questionnaires completed by the students, their mathematics teachers, and principals or other school administrators provided an abundance of contextual data within which to interpret the mathematics results.

The purpose of this report is to provide technical information about the 1992 Trial State Assessment in mathematics. It provides a description of the design for the Trial State Assessment and gives an overview of the steps involved in the implementation of the program from the planning stages through to the analysis and reporting of the data. The report describes in detail the development of the cognitive and background questions, the field procedures, the creation of the database for analysis (from receipt of the assessment materials through scanning, scoring, and creation of the database), and the methods and procedures for sampling, analysis, and reporting. It does not provide the results of the assessment—rather, it provides information on how those results were derived.

Educational Testing Service (ETS) was the contractor for the 1990 and 1992 NAEP programs, including the Trial State Assessment. ETS was responsible for overall management of the programs as well as for development of the overall design, the items and questionnaires, data analysis, and reporting. Westat, Inc., and National Computer Systems (NCS) were subcontractors to ETS. Westat was responsible for all aspects of sampling and of field operations, while NCS was responsible for printing, distribution, and receipt of all assessment materials, and for scanning and professional scoring.

This technical report provides supporting material for the series of reports that have been prepared for the 1992 Trial State Assessment Program in mathematics, including:

- A *State Report* for each participating jurisdiction that describes the mathematics proficiency of the fourth- and eighth-grade public-school students in that jurisdiction and relates their proficiency to contextual information about mathematics policies and instruction.

- The *NAEP 1992 Mathematics Report Card for the Nation and the States*, which provides data for all of the 44 jurisdictions that participated in the Trial State Assessment Program as well as the results from the 1992 national mathematics assessment.

- The *Executive Summary of the NAEP 1992 Mathematics Report Card for the Nation and the States*, providing the highlights of the *Mathematics Report Card*.

- The *Data Compendium from the NAEP 1992 Mathematics Assessment for the Nation and the States*, which includes tables of data relating performance on the mathematics assessment to a wide variety of demographic, perceptual, and experiential variables.

- *Interpreting NAEP Scales*, which describes past, present, and possible future methods of reporting and interpreting NAEP data. These include percent correct statistics, average percent correct, scale scores, scale anchoring, item mapping, and achievement levels.

- Two *Almanacs* for each jurisdiction, one for grade 4 and one for grade 8, that contain a detailed breakdown of the mathematics proficiency data according to the responses to the student, teacher, and school questionnaires for the population as a whole and for important subgroups of the population. There are five sections to each almanac:
  - *The Student Questionnaire Section* provides a breakdown of the proficiency data according to the students' responses to questions in the three student questionnaires included in the assessment booklets.
  - *The Teacher Questionnaire Section* provides a breakdown of the proficiency data according to the teachers' responses to questions in the mathematics teacher questionnaires.[2]
  - *The School Questionnaire Section* provides a breakdown of the proficiency data according to the principals' (or other administrators') responses to questions in the school characteristics and policies questionnaire.
  - *The Scale Section* provides a breakdown of selected questions from the questionnaires according to each of the scales measuring areas of mathematics in the assessment.[3]
  - *The Mathematics Item Section* provides the response data for each mathematics item in the assessment.

ORGANIZATION OF THE TECHNICAL REPORT

This chapter provides a description of the design for the Trial State Assessment in mathematics and gives an overview of the steps involved in implementing the program from the planning stage through the analysis and reporting of the data. The chapter summarizes the major components of the program with references to the appropriate chapters for more details. The organization of this chapter, and of the report, is as follows:

---

[2] Because both mathematics and reading were assessed at the fourth-grade level, the fourth-grade teacher questionnaire asked questions about mathematics and reading programs. The mathematics teachers of the students who participated in the mathematics assessment completed the mathematics questions and the reading teachers of the students in the reading assessment completed the reading questions. All teachers were asked to complete the questions about their educational background and training. For the mathematics assessment, only the data from the students' mathematics teachers are included.

[3] Scales were created for the content areas of Numbers and Operations; Measurement; Geometry; Data Analysis, Statistics, and Probability; and Algebra and Functions. An additional scale was created from items designed to measure estimation abilities.

- Section 1.2 provides an overview of the design of the Trial State Assessment Program.

- Section 1.3 summarizes the development of the mathematics objectives and the development and review of the items written to measure those objectives. Details are provided in Chapter 2.

- Section 1.4 discusses the assignment of the cognitive and background questions to assessment booklets and describes the focused-BIB spiral design. A complete description is provided in Chapter 2.

- Section 1.5 outlines the sampling design used for the Trial State Assessment Program. A fuller description is provided in Chapter 3.

- Section 1.6 summarizes the field administration procedures including securing school cooperation, training administrators, administering the assessment, and conducting quality control. Further details appear in Chapter 4.

- Section 1.7 describes the flow of the data from their receipt at National Computer Systems through data entry, professional scoring, and entry into the database for analysis. Chapters 5 and 6 provide a detailed description of the process.

- Section 1.8 provides an overview of the data obtained from the Trial State Assessment.

- Section 1.9 summarizes the procedures used to weight the data from the assessment and to obtain estimates of the sampling variability of subpopulation estimates. Chapter 7 provides a full description of the weighting and variance estimation procedures.

- Section 1.10 describes the initial analyses performed to verify the quality of the data in preparation for more refined analyses, with details given in Chapter 9.

- Section 1.11 describes the item response theory subscales and the overall mathematics composite that were created for the primary analysis of the Trial State Assessment data. Further discussion of the theory and philosophy of the scaling technology appears in Chapter 8 with details of the scaling process in Chapter 9.

- Section 1.12 provides an overview of the linking of the scaled results from the Trial State Assessment to those from the national mathematics assessment. Details of the linking process appear in Chapter 9.

- Section 1.13 describes the reporting of the assessment results, with further details supplied in Chapter 10.

- Appendices provide information about the participants in the objectives and item development process, a summary of the participation rates, a list of the conditioning variables, the IRT parameters for the mathematics items, the reporting subgroups, composite and derived common background and reporting variables, a description of the processes for anchoring the mathematics composite and for defining achievement levels on that scale, and a discussion of further analysis of 1990 data.

1.2 DESIGN OF THE TRIAL STATE ASSESSMENT IN MATHEMATICS

The major aspects of the design for the Trial State Assessment in mathematics included the following:

- Participation was voluntary.
- Only fourth- and eighth-grade students in public schools were assessed. Students in private or parochial schools were not included in the program. A representative sample of schools was selected in each participating jurisdiction, and students were randomly sampled within schools.
- Mathematics was assessed at the fourth- and eighth-grade levels.
- The mathematics items used in the Trial State Assessment were also used in the age 9/grade 4 and age 13/grade 8 national assessment and contained multiple-choice, short constructed-response, and extended constructed-response items. Some items required the use of calculators (four-function calculators at grade 4 and scientific calculators at grade 8), geometric shapes, and protractors/rulers. The total pool of mathematics items was divided into 13 15-minute blocks at each grade level. Each student in the Trial State Assessment also was assessed with a special block measuring estimation skills that was administered using an audiotape to pace students through the items.
- Background questionnaires given to the students, the students' mathematics teachers, and the principals or other school administrators provided for rich contextual information. The background questionnaires for the Trial State Assessment were identical to those used in the age 9/grade 4 and age 13/grade 8 national assessment.
- A complex form of matrix sampling called a balanced incomplete block (BIB) spiraling design was used. With BIB spiraling, students in an assessment session received different booklets, resulting in a more efficient sample. This design also reduced student burden and provided for greater mathematics content coverage than would have been possible had every student been administered the identical set of items.
- The assessment time for each student was approximately 71 minutes. Each assessed student was assigned a mathematics booklet that contained two five-minute background questionnaires, one 3-minute background questionnaire, and three of the thirteen 15-minute blocks containing mathematics items. Twenty-six different booklets were assembled. After the completion of this part of the assessment, each student was given another booklet containing the estimation questions. The estimation section took approximately 15 minutes.

- The assessments took place in the five-week period between February 3 and March 6, 1992. One-fourth of the schools in each state were assessed each week throughout the first four weeks; the fifth week was reserved for the scheduling of makeup sessions.

- Data collection, by law, was the responsibility of each participating jurisdiction.

- Security and uniform assessment administration were high priorities. Extensive training was conducted to assure that the assessment would be administered under standard, uniform procedures. Fifty percent of the assessment sessions were monitored by the contractor's staff.

1.3 DEVELOPMENT OF MATHEMATICS OBJECTIVES, ITEMS, AND BACKGROUND QUESTIONS

Similar to all previous NAEP assessments, the objectives for the Trial State Assessment in mathematics were developed through a broad-based consensus process managed by the Council of Chief State School Officers. Educators, scholars, and citizens, representative of many diverse constituencies and points of view, designed objectives for the mathematics assessment, proposing goals they believed students should achieve in the course of their education. After careful reviews of the objectives, assessment questions were developed that were appropriate to those objectives. Representatives from State Education Agencies provided extensive input throughout the entire development process.

The framework used for the 1992 mathematics assessment was the same as the one adopted for the 1990 assessment and was organized according to three mathematical abilities and five content areas. The mathematical abilities assessed were conceptual understanding, procedural knowledge, and problem solving. Additionally, students' abilities in estimation were assessed. Content was drawn primarily from elementary and secondary school mathematics up to, but not including, calculus. The content areas assessed were Numbers and Operations; Measurement; Geometry; Data Analysis, Statistics, and Probability; and Algebra and Functions.

The Trial State Assessment included multiple-choice, short constructed-response, and extended constructed-response items. All items underwent extensive reviews by specialists in mathematics, measurement, and bias/sensitivity, as well as reviews by representatives from State Education Agencies. In addition, the items were reviewed by representatives of the National Assessment Governing Board (NAGB) in accordance with NAGB's statutory responsibility for ensuring that all items selected for use in NAEP are free from racial, cultural, gender, or regional biases. The items were field tested on a representative group of students. Based on the results of the field test, items were revised or modified as necessary and then again reviewed for sensitivity, content, and editorial concerns. With the assistance of ETS/NAEP staff and outside reviewers, the mathematics Item Development Committee selected the items to include in the assessment.

The 1992 mathematics assessment at grade 8 was designed to estimate trends in performance for states that participated in both the 1990 and 1992 Trial State Assessments. To permit linking to the 1990 assessment, some of the items used in 1990 were used again in 1992.

Of the 175 fourth-grade items used in 1992, 57 (16 short constructed-response items and 41 multiple-choice items) had also been used in the 1990 program. Of the 205 eighth-grade items used in 1992, 76 (23 short constructed-response items and 53 multiple-choice items) had also been used in 1990. The rest of the items used in the 1992 program were newly created.

Chapter 2 includes specific details about developing the objectives and items for the Trial State Assessment. The details of the professional scoring process are given in Chapter 5.

1.4 ASSESSMENT INSTRUMENTS

The assembly of cognitive items into booklets and their subsequent assignment to assessed students was determined by a balanced incomplete block (BIB) design with spiraled administration. Details of the BIB design are provided in Chapter 2.

The student assessment booklets contained six sections and included both cognitive and noncognitive items. In addition to three sections of cognitive questions, each booklet included two 5-minute sets of general and mathematics background questions designed to gather contextual information about students, their experiences in mathematics, and their perceptions of the subject, and a section of questions designed to gather information about the students' levels of motivation in taking the assessment and their familiarity with the types of assessment questions they encountered.

In addition to the student assessment booklets, three other instruments provided data relating to the assessment—mathematics teacher questionnaires, school characteristics and policies questionnaires, and an excluded student questionnaire.

The teacher questionnaires were administered to the fourth- and eighth-grade mathematics teachers of the students participating in the assessment. At grade 4, the questionnaire consisted of three sections and took approximately 20 minutes to complete. The first section focused on teachers' background and experience. The second section focused on classroom information related to mathematics. The third section, which was completed only by those reading teachers of students in the fourth-grade Trial State Assessment in reading, focused on classroom information about reading. At grade 8, the mathematics teacher questionnaire consisted of two sections and also took approximately 20 minutes to complete. The first section focused on teachers' background and experience. The second section focused on mathematics classroom information.

The school characteristics and policies questionnaire was given to the principal or other administrator in each participating school and took about 15 minutes to complete. The questions asked about the principal's background and experience, school policies, programs, facilities, and the composition and background of the students and teachers.

The excluded student questionnaire was completed by the teachers of those students who were selected to participate in the Trial State Assessment sample but who were determined by the school to be ineligible to be assessed because they either had an Individualized Education Plan (IEP) and were not mainstreamed at least 50 percent of the time, or were categorized as Limited English Proficient (LEP). This questionnaire took approximately three minutes per student to complete and asked about the nature of the student's exclusion and the special programs in which the student participated.

1.5 THE SAMPLING DESIGN

The target population for the Trial State Assessment Program consisted of fourth- and eighth-grade students enrolled in public schools. The representative sample of students assessed in the Trial State Assessment came from about 125 public schools for grade 4 and 100 public schools for grade 8 in each jurisdiction, unless a jurisdiction had fewer than 125 schools with a fourth grade or fewer than 100 schools with an eighth grade, in which case all or almost all schools were asked to participate. The sample in each state was designed both to produce aggregate estimates for the state, and selected subpopulations (depending upon the size and distribution of the various subpopulations within the state), and also to enable comparisons to be made, at the state level, between administration with monitoring and without monitoring. The schools were stratified by urbanicity, percentage of Black and Hispanic students enrolled, and median household income.

At each grade level, 30 students selected from each school provided a sample size of approximately 3,000 students per state. The student sample size of 30 for each school was chosen to ensure at least 2,000 students participating from each state at each grade level for each subject area, allowing for school nonresponse, exclusion of students, inaccuracies in the measures of enrollment, and student absenteeism from the assessment.

The students within a school were sampled from lists of fourth- and eighth-grade students. The decisions to exclude students from the assessment were made by school personnel, as in the national assessment, and used the same specific criteria for exclusion (described in section 1.4) as in the national assessment. Each excluded student was carefully accounted for to estimate the percentage of the state population deemed unassessable and the reasons for exclusion.

Chapter 3 describes the various aspects of selecting the sample for the 1992 Trial State Assessment—the construction of the school frames, the stratification process, the updating of the school frame with new schools, the actual sample selection, and the sample selection for the field test.

1.6 FIELD ADMINISTRATION

The administration for the 1992 program and the 1991 field test involved a collaborative effort between staff in the participating states and schools and the NAEP contractors, especially Westat, the field administration contractor. The purpose of the field test conducted in 1991 was to try out the items and procedures for the 1992 program.

Each jurisdiction volunteering to participate in the 1991 field test and in the 1992 Trial State Assessment was asked to appoint a state coordinator who became the liaison between NAEP staff and the participating schools. At the local school level, an assessment administrator was responsible for preparing for and conducting the assessment session in one or more schools.

These individuals were usually school or district staff and were trained by Westat staff. In addition, Westat hired and trained a state supervisor for each state. The state supervisors were responsible for working with the state coordinators and overseeing assessment activities. Westat also hired and trained four to eight quality control monitors in each state to monitor 50 percent of the assessment sessions in 1992. During the field test, the state supervisors monitored all sessions.

Chapter 4 describes the procedures for obtaining cooperation from states and provides details about the field activities for both the field test and 1992 program. Chapter 4 also describes the planning and preparations for the actual administration of the assessment, the training and monitoring of the assessment sessions, and a description of the responsibilities and roles of the state coordinators, state supervisors, assessment administrators, and quality control monitors.

1.7 MATERIALS PROCESSING AND DATABASE CREATION

Upon completion of each assessment session, school personnel shipped the assessment booklets and forms from the field to NAEP subcontractor National Computer Systems for professional scoring, entry into computer files, and checking. Then the files were sent to Educational Testing Service for creation of the database. Careful checking assured that all data from the field were received. More than 498,000 booklets or questionnaires were received and processed for the mathematics assessment. The processing of these data is detailed in Chapter 5. That chapter also details the printing, distribution, receipt, processing, and final disposition of the 1992 Trial State Assessment materials.

The volume of collected data and the complexity of the Trial State Assessment processing design, with its spiraled distribution of booklets, as well as the concurrent administration of this assessment and the national assessments, required the development and implementation of flexible, innovatively designed processing programs and a sophisticated Process Control System. This system, which is described in Chapter 5, allowed an integration of data entry and workflow management systems, including carefully planned and delineated editing, quality control, and auditing procedures.

The data transcription and editing procedures are also described in Chapter 5. These procedures resulted in the generation of disk and tape files containing various assessment information, including the sampling weights required to make valid statistical inferences about the population from which the Trial State Assessment sample was drawn. Before any analysis could begin, the data from these files had to undergo a quality control process at ETS. The files were then merged into a comprehensive, integrated database. Chapter 6 describes the transcribed data files, the procedure of merging them, or bringing them together, to create the Trial State Assessment database, and the results of the quality control process.

1.8 THE TRIAL STATE ASSESSMENT DATA

Approximately 2,500 students at each grade were assessed within each state and the District of Columbia; apart from nonresponse, all fourth- and eighth-grade public-school students were assessed in Guam and the Virgin Islands.

The basic information collected from the Trial State Assessment consisted of the responses of the assessed students to the 158 mathematics exercises at grade 4 and 183 exercises at grade 8. To limit the assessment time for each student to about one hour, a variant of matrix sampling called BIB spiraling was used to assign a subset of the full exercise pool to each student. At each grade level, the set of items was divided into 13 unique blocks, each requiring 15 minutes for completion. Each assessed student received a booklet containing three of the 13 blocks according to a design that ensured that each block was administered to a representative sample of students within each jurisdiction. Following the administration of this booklet, each student was given a special booklet that contained the audiotaped estimation items. The data also included responses to the background questionnaires (described in section 1.4 and Chapter 2).

The national data to which the Trial State Assessment results were compared came from nationally representative samples of public-school students in the fourth and eighth grade. These samples were a part of the full 1992 national mathematics assessment in which nationally representative samples of students in public and private schools from three age cohorts were assessed: students who were either in the fourth grade or 9 years old; students who were either in the eighth grade or 13 years old; and students who were either in the twelfth grade or 17 years old.

The assessment instruments used in the Trial State Assessment were also used in the fourth- and eighth-grade national assessments and were administered using the identical procedures in both assessments. The time of testing for the state assessments (February 3 to March 6, 1992) occurred within the time of testing of the national assessment (January 6 to April 3, 1992). However, the state assessments differed from the national assessment in one important regard: Westat staff collected the data for the national assessment while, in accordance with the NAEP legislation, data collection activities for the Trial State Assessment were the responsibility of each participating jurisdiction. These activities included ensuring the participation of selected schools and students, assessing students according to standardized procedures, and observing procedures for test security. To provide quality control of the Trial State Assessment, a random half of the administrations within each state was monitored.

1.9 WEIGHTING AND VARIANCE ESTIMATION

The Trial State Assessment used a complex sample design to select the students to be assessed in each of the participating jurisdictions. The properties of a sample from a complex design are very different from those of a simple random sample in which every student in the target population has an equal chance of selection and in which the observations from different sampled students can be considered to be statistically independent of one another. The properties of the sample from the complex Trial State Assessment design were taken into account in the analysis of the assessment data.

One way that the properties of the sample design were taken into account was through the use of sampling weights which account for the fact that the probabilities of selection are not identical for all students. These weights also include adjustments for nonresponse of students and of schools. All population and subpopulation characteristics based on the Trial State Assessment data used the sampling weights in their estimation. Chapter 7 provides details on the computation of these weights.

In addition to deriving appropriate estimates of population characteristics, it is essential to obtain appropriate measures of the degree of uncertainty of those statistics. One component of uncertainty is a result of sampling variability, which measures the dependence of the results on the particular sample of students actually assessed. Because of the effects of cluster selection (first schools are selected and then students are selected within those schools), observations made on different students cannot be assumed to be independent of each other (and, in fact, are generally positively correlated). As a result, classical variance estimation formulae will produce incorrect results. Instead, a variance estimation procedure which does take the characteristics of the sample into account was used for all analyses. This procedure, called the jackknife variance estimator, is discussed in Chapter 7.

The jackknife variance estimator provides a reasonable measure of uncertainty for any statistic based on values observed without error. Statistics such as the average proportion of students correctly answering a given question meet this requirement but other statistics, based on estimates of student mathematics proficiency, such as the average mathematics proficiency of a subpopulation, do not. Because each student typically responds to relatively few items within a particular mathematics content area, there exists a nontrivial amount of imprecision in the measurement of the proficiency of any given student. This imprecision adds an additional component of variability to statistics based on estimates of individual proficiencies. The estimation of this component of variability is discussed in Chapter 8.

1.10 PRELIMINARY DATA ANALYSIS

Immediately after receipt from NCS of the machine-readable data tapes containing students' responses, all cognitive and noncognitive items were subjected to an extensive item analysis to assure that each item represented what it was purported to measure.

Each block of cognitive items was subjected to item analysis routines, which yielded, for each item, the number of respondents, the percentage of students who selected the correct response and each incorrect response, the percentage who omitted the item, the percentage who did not reach the item, and the correlation between the item score and the block score. In addition, the item-analysis program provided summary statistics for each block, including reliability (internal consistency). These kinds of analyses were used to check on the scoring of the items, to verify the appropriateness of the difficulty level of the items, and to check for speededness. The results also were reviewed by knowledgeable project staff in search of anomalies that might signal unusual results or errors in creating the database.

Tables of the weighted percentages of students choosing each of the possible responses to each cognitive and background item were created and distributed to each state and jurisdiction. Additional analyses comparing the data from the monitored sessions with that from the unmonitored sessions were conducted to determine the comparability of the assessment data from the two types of administrations. Finally, differential item functioning analyses were carried out to identify items that were differentially difficult for various subgroups and to reexamine such items with respect to their fairness and their appropriateness for inclusion in the scaling process. Further details of the preliminary analyses conducted on the data appear in Chapter 9.

1.11 SCALING THE ASSESSMENT ITEMS

The primary analysis and reporting of the results from the Trial State Assessment used item response theory (IRT) scale score models. Scaling models quantify a respondent's tendency to provide correct answers to the items contributing to a scale as a function of a parameter called proficiency that can be viewed as a summary measure of performance across all items entering into the scale. Three distinct IRT models were used for scaling: 1) 3-parameter logistic models for multiple choice items; 2) 2-parameter logistic models for simple constructed-response items that were scored correct or incorrect; and 3) generalized partial credit models for extended constructed response items that were scored on a multi-point scale. Chapter 8 provides an overview of the scaling models used, with further details on the application of these models provided in Chapter 9.

A series of scales were created for the Trial State Assessment to summarize students' mathematics performance. These scales were defined identically to those used for the scaling of the national NAEP fourth- and eighth-grade mathematics data. Five content area scales, based on the paradigm described in Chapter 2, were created to correspond to each of the following areas: Numbers and Operations; Measurement; Geometry; Data Analysis, Statistics, and Probability; and Algebra and Functions. An additional scale was created for other items designed to measure estimation abilities. Although the items comprising each scale were identical to those used for the national program, the item parameters for the Trial State Assessment scales were estimated from the combined data from all states and jurisdictions participating in the Trial State Assessment. Item parameter estimation was based on an item calibration sample consisting of an approximately 25 percent sample of all the available data. To ensure equal representation in the scaling process, each state and jurisdiction was equally represented in the item calibration sample, as were the monitored and unmonitored administrations from each state and jurisdiction. Chapter 9 provides further details about the item parameter estimation.

The fit of the IRT model to the observed data was examined within each scale by comparing the estimates of the empirical item characteristic functions with the theoretic curves. For binary-scored items, nonmodel-based estimates of the expected proportions of correct responses to each item for students with various levels of scale proficiency were compared with the fitted item response curve; for the extended constructed response items, the comparisons were based on the expected proportions of students with various levels of scale proficiency who achieved each score level. In general, the item level results were well fit by the scaling models.

Using the item parameter estimates, estimates of various population statistics were obtained for each jurisdiction in the Trial State Assessment. The NAEP methods use random draws ("plausible values") from estimated proficiency distributions for each student to compute population statistics. Plausible values are not optimal estimates of individual student proficiencies; instead, they serve as intermediate values to be used in estimating population characteristics. Under the assumptions of the scaling models, these population estimates will be consistent, in the sense that the estimates approach the model based population values as the sample size increases, which would not be the case for subpopulation estimates obtained by aggregating optimal estimates of individual proficiency. Chapter 8 provides further details on the computation and use of plausible values.

In addition to the plausible values for each scale, a composite of the content area scales was created at each grade as a measure of overall mathematics proficiency. This composite was a weighted average of the content area scale plausible values in which the weights were proportional to the relative importance assigned to each content area as specified in the mathematics objectives. Consistent with the mathematics framework, the weights used to define the composite were somewhat different at grades 4 and 8. The definitions of the composites for the Trial State Assessment program at grades 4 and 8 were identical to those used for the national fourth- and eighth-grade mathematics assessments.

1.12 LINKING THE TRIAL STATE RESULTS TO THE NATIONAL RESULTS

The results from the Trial State Assessment were linked to those from the national NAEP through linking functions determined by comparing the results for the aggregate of all students assessed in the Trial State Assessment at each of grades 4 and 8 with the results for students of the matching grade within the State Aggregate Comparison (SAC) subsample of the national NAEP. The SAC subsample of the national NAEP for a given grade is a representative sample of the population of all grade-eligible public-school students within the aggregate of the 41 participating states and the District of Columbia. Specifically, the grade 4 SAC subsample consists of all fourth-grade students in public schools in the states and the District of Columbia who were assessed in the national cross-sectional mathematics assessment. The grade 8 SAC subsample is equivalently defined for eighth-grade students who participated in the national assessment.

For each grade, a linear equating within each scale was used to link the results of the Trial State Assessment to the national NAEP. The adequacy of linear equating was evaluated by comparing, for each scale, the distribution of mathematics proficiency based on the aggregation of all assessed students at each grade from the participating states and the District of Columbia with the equivalent distribution based on the students in the SAC subsample for the matching grade. In the estimation of these distributions, the students were weighted to represent the target population of public-school students in the specified grade in the aggregation of the states and the District of Columbia (the students from Guam and the Virgin Islands were not included in the equating). If a linear equating is adequate, the distribution for the aggregate of states and the District of Columbia and that for the SAC subsample will have, to a close approximation, the same shape, in terms of the skewness, kurtosis, and higher moments of the distributions. The only differences in the distributions allowed by linear equating are in the means and variances. This was found to be the case.

The linking was accomplished for each grade and scale by matching the mean and standard deviation of the scale proficiencies across all students in each of grades 4 and 8 in the

Trial State Assessment (excluding Guam and the Virgin Islands) to the corresponding scale mean and standard deviation across all students in the matching grade SAC subsample. Further details of the linking are given in Chapter 9.

1.13 REPORTING THE TRIAL STATE ASSESSMENT RESULTS

Each state and jurisdiction that participated in the Trial State Assessment received multiple copies of a summary report providing the state's results with accompanying text and tables, and including national and regional comparisons. These reports were generated by a computerized report-generation system in which graphic designers, statisticians, data analysts, and report writers collaborated to develop shells of the reports in advance of the analysis. These prototype reports were provided to State Education Agency personnel for their reviews and comments. The results of the data analysis were then automatically incorporated into the reports that gave, in addition to tables and graphs of the results, interpretations of those results including indications of subpopulation comparisons of statistical and substantive significance.

Each report contained state-level estimates of mean proficiencies, both for the state as a whole and for categories of the key reporting variables: gender, race/ethnicity, level of parental education, and community type. Results were presented for each scale and for the overall mathematics composite. Results were also reported for a variety of other subpopulations based on variables derived from the student, teacher, and school questionnaires. Standard errors were included for all statistics.

A second report, the *NAEP 1992 Mathematics Report Card for the Nation and the States*, highlights key assessment results for the nation and summarizes results across the states and territories participating in the assessment. This report contains composite scale results (proficiency means, proportions at or above achievement levels, etc.) for the nation, each of the four regions of the country, and each jurisdiction participating in the Trial State Assessment, both overall and by the primary reporting variables. In addition, overall results are reported for each of the content area scales. For the jurisdictions that participated in both the 1990 and 1992 Trial State Assessments, reported results include trend comparisons to 1990.

The third type of summary report is entitled *Data Compendium from the NAEP 1992 Mathematics Assessment for the Nation and the States*. Like the *Report Card*, the *Compendium* reports results for the nation and for all of the states and territories participating in the Trial State Assessment. The *Compendium* contains most of the tables included in the *Report Card* plus additional tables that provide composite scale results for a large number of secondary reporting variables.

The fourth type of summary report is a five-section almanac. Three of the sections of the almanac (referred to as proficiency sections) present analyses based on responses to each of the questionnaires (student, mathematics teacher, and school) administered as part of the Trial State Assessment. The fourth section of the almanac, the scale section, reports proficiency means and associated standard errors for the five mathematics content-area scales and the estimation scale. Results in this section are also reported for the total group in each state, as well as for select subgroups of interest. The final section of the almanac, the "p-value" section, provides the total-group proportion of correct responses to each cognitive item included in the assessment.

The production of the state reports, *Mathematics Report Card, Data Compendium*, and the almanacs required a large number of decisions about a variety of data analysis and statistical issues. For example, because the demographic characteristics of the fourth- and eighth-grade public-school students vary widely by state, the proportions of students in the various categories of the race/ethnicity, parental education, and type of community variables varied by state. Chapter 10 documents the major conventions and statistical procedures used in generating the state reports, *Mathematics Report Card, Data Compendium*, and the almanacs. The chapter describes the rules, based on effect size and sample size considerations, that were used to establish whether a particular category contained sufficient data for reliable reporting of results for a particular state. Chapter 10 also describes the multiple comparison and effect size-based inferential rules that were used for evaluating the statistical and substantive significance of subpopulation comparisons.

To provide information about the generalizability of the results, a variety of information about participation rates was reported for each state and jurisdiction. This included the school participation rates, both in terms of the initially selected samples of schools and in terms of the finally achieved samples, including replacement schools. The student participation rates, the rates of students excluded due to Limited English Proficiency (LEP) and Individualized Education Plan (IEP) status, and the estimated proportions of assessed students who are classified as IEP or LEP were also reported by state.

Chapter 2

DEVELOPING THE MATHEMATICS OBJECTIVES, COGNITIVE ITEMS, BACKGROUND QUESTIONS, AND ASSESSMENT INSTRUMENTS

Stephen L. Koffler

Educational Testing Service

2.1 OVERVIEW

Similar to all previous NAEP assessments, the framework and objectives for the Trial State Assessment Program in mathematics were developed through a broad-based consensus process. Educators, scholars, and citizens, representative of many diverse constituencies and points of view, designed objectives for the mathematics assessment, proposing goals they believed students should achieve in the course of their education. The framework and objectives were initially developed for the 1990 mathematics assessment. They were used in both the 1990 Trial State Assessment and the national NAEP programs and were used again for both programs in 1992. In both years, the Trial State Assessment was a subset of the national mathematics assessment. The same objectives and instruments were used in both. After careful reviews of the objectives, assessment items were developed that were appropriate to those objectives. All items underwent extensive reviews by specialists in mathematics, measurement, and bias/sensitivity, as well as reviews by state representatives.

The objectives and item development efforts were governed by four major considerations:

- As specified in the 1988 NAEP legislation, the objectives had to be developed through a consensus process involving subject-matter experts, school administrators, teachers, and parents, and the items had to be carefully reviewed for potential bias.
- As outlined in the ETS proposal for the administration of the NAEP contract, the development of the items had to be guided by a Mathematics Item Development Panel.
- As described in the ETS Standards for Quality and Fairness (ETS, 1987), all materials developed at ETS had to be in compliance with specified procedures.
- As per federal regulations, all cognitive and background items had to be submitted to a federal clearance process.

This chapter includes specific details about developing the objectives and items for the 1992 mathematics assessment. The chapter also describes the instruments—the student assessment booklets (and the manner in which the items were organized into blocks to create the booklets), mathematics teacher questionnaires, school characteristics and policies questionnaires, and excluded student questionnaire. Many committees worked on the development of the framework, objectives, and items for the Trial State Assessment. A list of the committees and consultants participating in the development process for the 1992 mathematics assessment is included in Appendix A.

2.2 CONTEXT FOR PLANNING THE 1992 MATHEMATICS ASSESSMENT[1]

Anticipating the 1988 legislation that authorized the Trial State Assessment, in mid-1987 the federal government arranged for a special grant from the National Science Foundation and the U.S. Department of Education to the Council of Chief State School Officers to prepare the framework and objectives and make recommendations about reporting for the Trial State Assessment Program in mathematics.

The Council of Chief State School Officers established the National Assessment Planning Project to oversee their work for the Trial State Assessment. The National Assessment Planning Project, whose members included policymakers, practitioners, and citizens nominated by 18 national organizations, had two primary purposes—to recommend objectives for the 1990 Trial State Assessment in eighth-grade mathematics and to make suggestions for reporting the results from that program. However, because the 1990 objectives had to be coordinated across the three grades, the objectives developed by the Project governed the entire NAEP mathematics assessment, including the national assessment at grades 4, 8, and 12 as well as the Trial State Assessment at grade 8. This was also true for 1992.

2.3 ASSESSMENT DESIGN PRINCIPLES

The Council of Chief State School Officers created a Mathematics Objectives Committee to recommend objectives for the assessment. The Committee consisted of a teacher, a school administrator, mathematics education specialists from various states, mathematicians, parents, and citizens.

Two principles emerged during the discussions of the Mathematics Objectives Committee and became the basis for structuring the framework and objectives for the assessment. The first principle was that a national assessment, designed to provide state-level comparisons, should not measure only those topics and skills already included in the objectives of all states nor be geared to the *least common denominator* of student preparation. The second principle was that the assessment should not be used to steer instruction toward one particular pedagogical or philosophical viewpoint to the exclusion of others that are widely held.

---

[1] For more details see *Mathematics Objectives, 1990 Assessment* (National Assessment of Educational Progress, 1988).

The objectives development was also guided by several other considerations: that the assessment should 1) reflect many of the states' curricular emphases and objectives; 2) reflect what various scholars, practitioners, and interested citizens believe should be included in the curriculum; and 3) maintain some of the content of prior assessments to permit reporting of trends in performance. Accordingly, the committee gave attention to several frames of reference:

- states' goals and concerns, as reflected through analyses of state mathematics curriculum guides and the recommendations of state mathematics specialists;
- a report on "Issues in the Field," based on telephone interviews with leading mathematics educators, and a draft assessment framework provided by a subcommittee of the Mathematics Objectives Committee;
- the draft of the *Curriculum and Evaluation Standards for School Mathematics*, developed by the National Council of Teachers of Mathematics through intensive work by leading mathematics educators in the United States (NCTM, 1987); and
- the design of the 1986 mathematics assessment (NAEP, 1987). The framework for the 1986 NAEP mathematics assessment had 35 cells—seven content and five process areas. Because there were so many cells, the weightings assigned to some of the cells in the 1986 framework did not result in a sufficient number of items to provide reliable measures of students' knowledge and skills. As a result, it was decided that the outline or matrix guiding the development of the 1990 mathematics assessment had to be simplified—rather than having a large number of cells, necessary complexity could be reflected through the designation of specific abilities and topics in each content area.

2.4 ASSESSMENT DEVELOPMENT PROCESS

The Mathematics Objectives Committee developed a draft framework, set of objectives, and set of sample items, which were distributed to the mathematics supervisor in each of the 50 State Education Agencies. These supervisors convened a panel that reviewed the draft and returned comments and suggestions to the project staff. Copies of the draft were also sent to 25 mathematics educators and scholars for review. The Mathematics Objectives Committee incorporated the recommendations made and formulated their final recommendations, which were approved by the National Assessment Planning Project Steering Committee.

The framework and objectives were then submitted to the National Center for Education Statistics, which forwarded them for review to the Assessment Policy Committee, a panel that advised on NAEP policy at that time. The Assessment Policy Committee approved the objectives with minor provisions about the feasibility of full implementation.[3] The framework and objectives were refined by NAEP's Mathematics Item Development Panel, reviewed by the Task Force on State Comparisons, and resubmitted to the National Center for Education Statistics for adoption.

---

[3] This action was contained in a statement issued by the Assessment Policy Committee's Executive Committee on April 29, 1988. The recommendations were ratified by the full committee on June 18, 1988, with two stipulations: that the objectives be so weighted as to permit reporting on trends in performance; and, with regard to the use of calculator-

2.5 MATHEMATICS FRAMEWORK

The framework adopted for the 1990 mathematics assessment (and therefore also for the 1992 mathematics assessment) is organized according to three mathematical abilities and five content areas. The mathematical abilities assessed were conceptual understanding, procedural knowledge, and problem solving. Content was drawn primarily from elementary and secondary school mathematics up to, but not including, calculus. The content areas assessed were numbers and operations; measurement; geometry; data analysis, statistics, and probability; and algebra and functions.

The assignment of the percentages of assessment items to be devoted to each mathematical ability and content area was an important component of the framework development because such weighting reflects the importance or value given to each area at each grade level. The National Assessment Planning Project wanted to create an assessment that would be forward-thinking and could lead instruction; thus, they decided to give more emphasis than in previous assessments to problem solving, geometry, and algebra and functions, and less to numbers and operations.

The distribution of items by mathematical ability and mathematical content area for each grade as defined in the framework is provided in Table 2-1 and Table 2-2.

Table 2-1

NAEP Mathematics Framework:
Percentage Distribution of Items by Grade and Ability

	Grade	
Mathematical Ability	4	8
Conceptual Understanding	40%	40%
Procedural Knowledge	30%	30%
Problem Solving	30%	30%

---

active items and open response questions, that the assessment be developed within the resources available for its administration.

Table 2-2

NAEP Mathematics Framework:
Percentage Distribution of Items by Grade and Content Area

Mathematical Content Area	Grade	
	4	8
Numbers and Operations	45%	30%
Measurement	20%	15%
Geometry	15%	20%
Data Analysis, Statistics, and Probability	10%	15%
Algebra and Functions	10%	20%

2.6 COGNITIVE ITEM DEVELOPMENT

The 1992 mathematics assessment was designed to estimate trends from 1990 in national performance at all three grade levels and at grade 8 for states that participated in both the 1990 and 1992 Trial State Assessments.

Both the 1990 and 1992 Trial State Assessments in mathematics included constructed-response and multiple-choice items. The 1992 assessment relied much more on constructed-response items than did the 1990 assessment. In addition to short constructed-response items, the 1992 assessment included extended constructed-response questions. The extended constructed-response mathematics items, which were used for the first time in the 1992 assessment, called for the student to work through a complex problem requiring about five minutes to complete, and were scored on a 0-4 scale.

All of the constructed-response items were designed to provide an extended view of students' mathematical knowledge and skills. Building on recommendations from the report of the Council of Chief State School Officers, the NAEP Mathematics Item Development Panel suggested that constructed-response items be used to assess objectives in the framework that are best measured using such types of items (e.g., the ability to articulate mathematical ideas, draw figures, or generalize function relationships). About half of the constructed-response questions required short answers; the other half, including the extended constructed-response questions, required the ability to formulate and demonstrate more detailed problem-solving skills.

To permit linking to the 1990 assessment, some of the items used in 1990 were used again in 1992. At grade 4, 57 items that were used in the 1990 program were carried forward to the 1992 program (16 short constructed-response items and 41 multiple-choice items). At grade 8, 76 items were used again (23 short constructed-response items and 53 multiple-choice items) and at grade 12, 80 items were reused (24 short constructed-response and 56 multiple-choice items). The rest of the items used in the 1992 program were newly created. In total, the 1992 assessment included many more items than did the 1990 assessment.

Similar to the development of the items for the 1990 assessment, a carefully developed and proven series of steps were used to create multiple-choice and short constructed-response items for 1992 that reflected the objectives.

1) The Mathematics Item Development Panel provided guidance to the NAEP staff about how the objectives could be measured given the constraints of resources and the feasibility of measurement technology. The Panel made recommendations about priorities for the assessment and types of items to be developed.

2) The content and ability classifications of the items from the 1990 assessment that were released to the public were determined so that new items could be developed to meet those specifications. This was necessary so that the overall proportion of items for 1992 would continue to meet the proportions called for in the framework.

3) Item writers, both within and outside ETS, with subject-matter expertise and skills and experience in creating items according to specifications, wrote assessment items.

4) The items were reviewed and revised by NAEP/ETS staff and external reviewers. In addition, the items were reviewed by representatives of the National Assessment Governing Board in accordance with the board's statutory responsibility for ensuring that all items selected for use in NAEP are free from racial, cultural, gender, or regional biases.

5) Representatives from the State Education Agencies met and reviewed all items and background questionnaires (see section 2.9 for a discussion of the background questionnaires).

6) Language editing and sensitivity reviews were conducted according to ETS quality control procedures.

7) Field test materials were prepared, including the materials necessary to secure Office of Management and Budget clearance.

The field tests for the multiple-choice and short constructed-response items were conducted in February 1991 in 22 states, the District of Columbia, and the Virgin Islands. The intent of the field test was to try out the items and procedures and to give the states and the contractors practice and experience with the proposed materials and procedures. About 500-600 responses were obtained for each mathematics item in the field test.

The field test data were scored and analyzed in preparation for meetings with the Mathematics Item Development Panel and the Background Panel. Using item analysis procedures, which provide a variety of statistics about each item in the field test (including p-values, biserial correlations, and item response theory plots), committee members, ETS test development staff, and NAEP/ETS staff reviewed the materials to determine

- the most appropriate items for use in the 1992 assessment in accordance with content specifications (that they met the content and ability specifications in the framework) and statistical specifications (that their biserial correlation was not less than 0.20);

- the need for revisions to items that lacked clarity or had ineffective item statistics; and
- appropriate timing for assessment items.

Once the pool of newly created items was established, the items were assembled into nine different "blocks" (15-minute sections established according to statistical guidelines developed at the beginning of the process).[3] The new blocks were assembled taking into account the speededness data from the field test and the fact that extended constructed-response items would be included in certain of the blocks.

The development of the extended constructed-response items was on a somewhat different set of timelines than the multiple-choice and short constructed-response items. A committee of mathematics educators from elementary and secondary schools, colleges, and state education agencies met early in 1991, and worked with ETS/NAEP mathematics test development staff to develop extended constructed-response items and scoring guides. These items were carefully reviewed according to the procedures required by the ETS *Standards for Quality and Fairness* (ETS, 1987), including content and sensitivity reviews.

Twelve items at each grade level were field tested in May 1991 in urban, suburban, and rural school districts in New Jersey and Pennsylvania. Each student was administered two extended constructed-response items and each item was given to approximately 50-100 students. ETS/NAEP mathematics test development staff scored the extended constructed-response items at a special two-day scoring session. Based on the distribution of scores and on the content specifications, the final set of extended constructed-response items was selected by ETS/NAEP mathematics test development staff and reviewed by the Mathematics Item Development Committee. These items were included as the last item in the appropriate blocks.

Once the total set of items had been selected and assembled into blocks, all items and blocks were reviewed again by ETS/NAEP staff for content, measurement, and sensitivity concerns. In addition, another meeting of representatives from State Education Agencies was convened to review the field test results and final set of items. The federal clearance process was initiated in August 1991 with the submission of materials to the National Center for Education Statistics. Revisions were made in accordance with changes required by the National Center for Education Statistics and NAGB and the final clearance package was approved in September 1991.

The overall pool of items (new and trend items) for the 1992 Trial State Assessment consisted of 158 items in grade 4 (54 short and 5 extended constructed-response items, and 99 multiple-choice items) and 183 items in grade 8 (59 short and 6 extended constructed-response items, and 118 multiple-choice items). For each grade, about 40 percent of the assessment time (35 percent of the items) was devoted to short and extended constructed-response questions.

---

[3] In total, there were 13 blocks at each grade level, nine newly created blocks and four trend blocks that had been used in the 1990 mathematics assessment.

Table 2-3 (grade 4) and Table 2-4 (grade 8) provide the number and percentage of items for each content and ability group for each grade. These items were also used in the national mathematics assessment.

Students participating in the 1992 Trial State Assessment were also administered a special set of items measuring estimation skills. The estimation items were presented to students in a separate booklet, accompanied by a paced audiotape, and were the same items that were included in the special studies that were part of the 1990 and 1992 national NAEP mathematics assessment. The estimation items assessed students' skills in making estimates appropriate to a wide variety of situations. The information from these items supplemented the data for the content areas of numbers and operations and measurement. At grade four, there were 20 items measuring estimation skills and at grade 8, there were 22 items.

Table 2-5 provides the number of questions that were included in the assessment at each grade level and Table 2-6 provides the number of questions at each grade level for the 1992 assessment that were contributed by the 1990 trend blocks (except for the estimation block which was an intact block administered as part of a special study in 1990 in the national assessment).

2.7 BLOCK DESIGN

The assessment included 13 different 15-minute blocks of multiple-choice and constructed-response items at each grade level. At each grade level, four blocks used in 1990 were retained for reassessment in 1992, including one calculator block and the protractor/ruler block; nine blocks were newly developed. Of the 13 blocks at each grade level:

- Three blocks included items designed to be answered using a calculator. For the grade 4 calculator blocks, students were provided with a four-function calculator, while at grade 8 students were provided with a scientific calculator.
- One block contained items requiring the use of a ruler at grade 4 and protractor/ruler at grade 8.
- One block contained questions about geometry for which students were given a set of geometric shapes to use.
- Five blocks at grade 4 and six blocks at grade 8 included extended constructed-response questions.

Table 2-7 (grade 4) and Table 2-8 (grade 8) provide the composition of each block of items administered in the Trial State Assessment Program.

Table 2-3

Number and Percentage of Questions for the 1990 and 1992 NAEP Assessments According to the Mathematics Framework, Grade 4

Mathematical Ability	Mathematical Content Area					
	Numbers and Operations	Measurement	Geometry	Data Analysis, Statistics, and Probability	Algebra and Functions	TOTAL
Conceptual Understanding	1990: 18 1992: 21	1990: 6 1992: 14	1990: 8 1992: 14	1990: 4 1992: 8	1990: 6 1992: 7	Goal: 40% 1990: 39% (42) 1992: 40% (64)
Procedural Knowledge	1990: 16 1992: 19	1990: 9 1992: 7	1990: 1 1992: 1	1990: 3 1992: 3	1990: 2 1992: 1	Goal: 30% 1990: 28% (31) 1992: 20% (31)
Problem Solving	1990: 18 1992: 23	1990: 6 1992: 10	1990: 5 1992: 12	1990: 1 1992: 9	1990: 6 1992: 9	Goal: 30% 1990: 33% (36) 1992: 40% (63)
TOTAL	Goal: 45% 1990: 48% (52) 1992: 40% (63)	Goal: 20% 1990: 19% (21) 1992: 20% (31)	Goal: 15% 1990: 13% (14) 1992: 17% (27)	Goal: 10% 1990: 7% (8) 1992: 12% (20)	Goal: 10% 1990: 13% (14) 1992: 11% (17)	Total Items: 1990: 109 1992: 158

Table 2-4

Number and Percentage of Questions for the 1990 and 1992 NAEP Assessments
According to the Mathematics Framework, Grade 8

Mathematical Ability	Mathematical Content Area					
	Numbers and Operations	Measurement	Geometry	Data Analysis, Statistics, and Probability	Algebra and Functions	TOTAL
Conceptual Understanding	1990: 18 1992: 23	1990: 7 1992: 8	1990: 13 1992: 14	1990: 9 1992: 11	1990: 12 1992: 11	Goal: 40% 1990: 43% (59) 1992: 37% (67)
Procedural Knowledge	1990: 15 1992: 18	1990: 9 1992: 9	1990: 4 1992: 5	1990: 5 1992: 6	1990: 8 1992: 7	Goal: 30% 1990: 30% (41) 1992: 24% (45)
Problem Solving	1990: 12 1992: 17	1990: 5 1992: 15	1990: 9 1992: 17	1990: 5 1992: 11	1990: 6 1992: 11	Goal: 30% 1990: 27% (37) 1992: 39% (71)
TOTAL	Goal: 30% 1990: 33% (45) 1992: 32% (58)	Goal: 15% 1990: 15% (21) 1992: 17% (32)	Goal: 20% 1990: 19% (26) 1992: 20% (36)	Goal: 15% 1990: 14% (19) 1992: 15% (28)	Goal: 20% 1990: 19% (26) 1992: 16% (29)	Total Items: 1990: 137 1992: 183

Table 2-5

Total Number of Items in the 1992 Trial State Assessment in Mathematics

<table>
<tr><th rowspan="2">Use of Questions</th><th colspan="2">Number of Questions at Each Grade</th></tr>
<tr><th>Grade 4</th><th>Grade 8</th></tr>
<tr><td>Grade 4 and Not Grade 8</td><td>78</td><td></td></tr>
<tr><td>Grade 8 and Not Grade 4</td><td></td><td>103</td></tr>
<tr><td>Grades 4 and 8</td><td colspan="2">80</td></tr>
<tr><td>Total per grade</td><td>158</td><td>183</td></tr>
<tr><td>Number short constructed-response</td><td>54</td><td>59</td></tr>
<tr><td>Number extended constructed-response</td><td>5</td><td>6</td></tr>
<tr><td>Number multiple-choice</td><td>99</td><td>118</td></tr>
</table>

Table 2-6

Number of Items Contributed by 1990 Blocks in the 1992 Trial State Assessment in Mathematics

<table>
<tr><th rowspan="2">Use of Questions</th><th colspan="2">Number of Questions at Each Grade</th></tr>
<tr><th>Grade 4</th><th>Grade 8</th></tr>
<tr><td>Grade 4 and Not Grade 8</td><td>15</td><td></td></tr>
<tr><td>Grade 8 and Not Grade 4</td><td></td><td>34</td></tr>
<tr><td>Grades 4 and 8</td><td colspan="2">42</td></tr>
<tr><td>Total per grade</td><td>57</td><td>76</td></tr>
<tr><td>Number short constructed-response</td><td>16</td><td>23</td></tr>
<tr><td>Number multiple-choice</td><td>41</td><td>53</td></tr>
</table>

Table 2-7

Cognitive and Noncognitive Block Composition, Grade 4

Block	Type	Total Number of Items	Number of Multiple-Choice Items	Number of Constructed-response Items		Booklets Containing Block
				Short	Extended	
B1	Common Background	20	20	0	0	1 - 26
M2	Mathematics Background	18	18	0	0	1 - 26
MB	Motivation Background	5	5	0	0	1 - 26
M3	Mathematics Cognitive	13	9	4	0	1, 13, 20
M4	Mathematics Cognitive (Trend)	14	14	0	0	1, 2, 21
M5	Mathematics Cognitive (Trend/Ruler)	17	13	4	0	2, 3, 22
M6	Mathematics Cognitive (Trend)	11	0	11	0	3, 4, 23
M7	Mathematics Cognitive	10	6	3	1	4, 5, 24
M8	Mathematics Cognitive (Trend/Calculator)	15	14	1	0	5, 6, 25
M9	Mathematics Cognitive	12	9	2	1	6, 7, 26
M10	Mathematics Cognitive (Manipulatives)	6	0	6	0	4, 7, 21
M11	Mathematics Cognitive	16	11	5	0	5, 8, 22
M12	Mathematics Cognitive (Calculator)	12	5	7	0	6, 9, 23
M13	Mathematics Cognitive	12	6	5	1	7, 10, 24
M14	Mathematics Cognitive (Calculator)	10	6	3	1	8, 11, 25
M15	Mathematics Cognitive	10	6	3	1	9, 12, 26

Table 2-8

Cognitive and Noncognitive Block Information, Grade 8

Block	Type	Total Number of Items	Number of Multiple-Choice Items	Number of Constructed-response Items		Booklets Containing Block
				Short	Extended	
B1	Common Background	22	22	0	0	1 - 26
M2	Mathematics Background	23	23	0	0	1 - 26
MB	Motivation Background	5	5	0	0	1 - 26
M3	Mathematics Cognitive	13	9	3	1	1, 13, 20
M4	Mathematics Cognitive (Trend)	21	21	0	0	1, 2, 21
M5	Mathematics Cognitive (Trend/Ruler)	21	16	5	0	2, 3, 22
M6	Mathematics Cognitive (Trend)	16	0	16	0	3, 4, 23
M7	Mathematics Cognitive	13	7	5	1	4, 5, 24
M8	Mathematics Cognitive (Trend/Calculator)	18	16	2	0	5, 6, 25
M9	Mathematics Cognitive	9	5	3	1	6, 7, 26
M10	Mathematics Cognitive (Manipulatives)	7	0	7	0	4, 7, 21
M11	Mathematics Cognitive	19	13	6	0	5, 8, 22
M12	Mathematics Cognitive (Calculator)	9	6	2	1	6, 9, 23
M13	Mathematics Cognitive	11	6	4	1	7, 10, 24
M14	Mathematics Cognitive (Calculator)	9	6	2	1	8, 11, 25
M15	Mathematics Cognitive	17	13	4	0	9, 12, 26

BEST COPY AVAILABLE

2.8 STUDENT ASSESSMENT BOOKLETS

The assembly of mathematics items into booklets and their subsequent assignment to assessed students was determined by a *balanced incomplete block* (BIB) design with *spiraled* administration.

The first step in implementing BIB spiraling required dividing the total pool of mathematics items into blocks designed to take 15 minutes to complete. These blocks were then assembled into booklets containing two 5-minute background sections, three blocks of mathematics items according to a partially balanced incomplete block design, and an additional 1-minute background section. Thus, the assessment time for each of these student booklets was approximately 56 minutes. Following the completion of the assessment booklet, all students were given another booklet that contained the paced audiotape block of estimation items which took about 15 minutes. Thus, the overall assessment time for each student was approximately 71 minutes.

The mathematics blocks were assigned to booklets in such a way that each block appeared in the same number of booklets and every pair of blocks appeared together in exactly one booklet. This is the *balanced* part of the balanced incomplete block design. It is an *incomplete* block design because no booklet contained all items and hence there is *incomplete* data for each assessed student.

The BIB design for the 1992 national mathematics assessment (and, therefore, for the Trial State Assessment) was *focused*—each block was paired with every other mathematics block but not with blocks from other subject areas. The *focused*-BIB design also balances the order of presentation of the blocks of items—every block appears as the first cognitive block in one booklet, as the second block in another booklet, and as the third block in a third booklet.

The focused-BIB design used at each grade level in 1992 required that 13 blocks of mathematics items be assembled into 26 booklets. The assessment booklets were then *spiraled* and bundled. Spiraling involves interleaving the booklets in a systematic sequence so that each booklet appears an appropriate number of times in the sample. The bundles were designed so that each booklet would appear equally often in each position in a bundle.

The final step in the BIB-spiraling procedure is the assigning of the booklets to the assessed students. The students within an assessment session were assigned booklets in the order in which the booklets were bundled. Thus, students in an assessment session received different booklets, and only several students in the session received the same booklet. In the Trial State Assessment BIB-spiral design, representative and randomly equivalent samples of between approximately 200 and 700 students for each jurisdiction responded to each item. Table 2-9 provides the total number of booklets, cognitive blocks, and noncognitive blocks used for the program. Table 2-9 also provides the details of the focused-BIB design that was used with 13 blocks and 26 booklets.

Table 2-9

Booklet Content at Each Grade Level

Booklet Number	Common Background Block	Mathematics Background Block	Cognitive Blocks			Motivation Background Block
1	B1	M2	M3	M4	M7	MB
2	B1	M2	M4	M5	M8	MB
3	B1	M2	M5	M6	M9	MB
4	B1	M2	M6	M7	M10	MB
5	B1	M2	M7	M8	M11	MB
6	B1	M2	M8	M9	M12	MB
7	B1	M2	M9	M10	M13	MB
8	B1	M2	M10	M11	M14	MB
9	B1	M2	M11	M12	M15	MB
10	B1	M2	M12	M13	M3	MB
11	B1	M2	M13	M14	M4	MB
12	B1	M2	M14	M15	M5	MB
13	B1	M2	M15	M3	M6	MB
14	B1	M2	M3	M5	M10	MB
15	B1	M2	M4	M6	M11	MB
16	B1	M2	M5	M7	M12	MB
17	B1	M2	M6	M8	M13	MB
18	B1	M2	M7	M9	M14	MB
19	B1	M2	M8	M10	M15	MB
20	B1	M2	M9	M11	M3	MB
21	B1	M2	M10	M12	M4	MB
22	B1	M2	M11	M13	M5	MB
23	B1	M2	M12	M14	M6	MB
24	B1	M2	M13	M15	M7	MB
25	B1	M2	M14	M3	M8	MB
26	B1	M2	M15	M4	M9	MB

Blocks M4, M5, M6, and M8 are trend blocks from 1990
Block M5 requires a protractor/ruler (grade 8) or ruler (grade 4)
Blocks M8, M12, M14 require a calculator
Block M10 requires geometric shapes/manipulatives

2.9 QUESTIONNAIRES

As part of the Trial State Assessment (as well as the national assessment), a series of background questionnaires was administered to students, teachers, and school administrators. Similar to the development of the cognitive items, the development of the policy issues and questionnaire items was an iterative process that involved staff work, field testing, and review by external advisory groups and the federal government. A Background Panel drafted a set of policy issues and made recommendations regarding the design of the questions. They were particularly interested in capitalizing on the unique properties of NAEP and not duplicating other surveys (e.g., The National Survey of Public and Private School Teachers and

Administrators, The School and Staffing Study, and The National Educational Longitudinal Study).

The Panel recommended a focused study that addressed the relationship between student achievement and instructional practices. For the 1992 assessment, the framework focused on five educational areas: instructional content, instructional practices and experiences, teacher characteristics, school conditions and context, and conditions beyond school (i.e., home support, out-of-school activities, and attitudes (NAEP, 1992). The items were written by ETS staff and reviewed by the Background Panel, representatives from State Education Agencies, the National Center for Education Statistics, and the Office of Management and Budget. The questionnaires were assembled into questionnaires and underwent internal ETS review procedures to ensure fairness and quality. They were field tested as part of the February 1991 field test and reviewed again by the Background Panel, representatives from State Education Agencies, the National Center for Education Statistics, and the Office of Management and Budget.

2.9.1 Student Questionnaires

In addition to the cognitive questions, the 1992 Trial State Assessment included two five-minute sets of general and mathematics background questions designed to gather contextual information about students, their experiences in mathematics, and their perceptions of the subject, and a one-minute set of background questions about the students' motivation regarding the assessment. In many cases the questions used were continued from prior assessments, especially from the 1990 assessment in order to measure change between 1990 and 1992.

The **student demographics (common core) questionnaire** (20 questions at grade 4 and 22 questions at grade 8) included questions about race/ethnicity, language spoken in the home, mother's and father's level of education, reading materials in the home, television watching, homework, and which parents live at home. This questionnaire was the first section in every booklet.

Three categories of information were represented in the second five-minute **student mathematics questionnaire** (18 questions at grade 4 and 23 questions at grade 8): time spent on task and mathematics coursework, the nature of students' mathematics instruction, and students' enjoyment of and confidence in their abilities in mathematics and their perceptions of the usefulness of the discipline to their present and future lives. This questionnaire was the second section in every booklet.

The **motivation questionnaire** (5 questions at each grade level) asked the students questions about their perceptions of the difficulty of the assessment, and of how well they did on the assessment, and their motivation to do well on the assessment. This questionnaire was the last section in every booklet.

2.9.2 Teacher, School, and Excluded Student Questionnaires

To supplement the information on instruction reported by students, the mathematics teachers of the fourth- and eighth-grade students participating in the Trial State Assessment were asked to complete a mathematics teacher questionnaire about their instructional practices, teaching backgrounds, and characteristics. The teacher questionnaires contained two parts.[4] **The Teacher Questionnaire, Part I: Background and Training** (23 questions at grade 4 and 32 questions at grade 8) included questions pertaining to gender, race/ethnicity, years of teaching experience, certification, degrees, major and minor fields study, coursework in education, coursework in subject area, in-service training, extent of control over classroom, instruction, and curriculum, and availability of resources for their classroom. **The Teacher Questionnaire, Part II: Class by Class Mathematics Information** (40 questions at grade 4 and 42 questions at grade 8) pertained to the procedures the teacher uses for *each class* containing an assessed student and included questions on the ability level of students in the class, whether students were assigned to the class by ability level, time on task, homework assignments, frequency of instructional activities used in class, instructional emphasis given to the topics and skills covered in the assessment, and use of particular resources.

A **School Characteristics and Policies Questionnaire** was given to the principal or other administrator of each school that participated in the Trial State Assessment Program. This questionnaire (77 questions at both grades 4 and 8) included questions about background and characteristics of school principals, length of school day and year, school enrollment, absenteeism, drop-out rates, policies about tracking, curriculum, testing practices and use, special priorities and school-wide programs, availability of resources, special services, community services, policies for parental involvement, and school-wide problems.

The **Excluded Student Questionnaire** was completed by the teachers of those students who were selected to participate in the Trial State Assessment sample but who were determined by the school to be ineligible to be assessed because they either had an Individualized Education Plan (IEP) and were not mainstreamed at least 50 percent of the time, or were categorized as Limited English Proficient (LEP). This questionnaire asked about the nature of the student's exclusion and the special programs in which the student participated.

Schools were permitted to exclude certain students from the assessment. The same exclusion criteria and rules used in the national assessment were also applied to the Trial State Assessment. Although the intent was to assess all sampled students, students who were identified by school staff as not capable of participating meaningfully were excluded. The NAEP guidelines for exclusion are intended to assure uniformity of exclusion criteria from school to school as well as from state to state.

---

[4] Because the Trial State Assessment at grade four included both mathematics and reading, the fourth grade teacher questionnaire contained three sections. The first asked about the teachers' background and training, the second asked about classroom information for the mathematics teachers of the students involved in the mathematics assessment, and the third asked about classroom information for the reading teachers of the students involved in the reading assessment. Mathematics teachers of students participating in the mathematics assessment were asked to complete parts one and two, only.

Chapter 3

SAMPLE DESIGN AND SELECTION

Leyla K. Mohadjer, Keith F. Rust, Valerija Smith, and Jacqueline Severynse

Westat, Inc.

3.1 Introduction and Overview

The 1992 Trial State Assessment Program included assessments in eighth-grade mathematics, fourth-grade mathematics, and fourth-grade reading. Three representative samples of public-school students were drawn in each participating state or territory. Each sample was designed to produce aggregate estimates as well as estimates for various subpopulations with approximately equal precision for the participating states. The sample for the eighth-grade assessment of mathematics consisted of about 2,500 eighth-grade students from about 100 public schools in each state or territory. Similarly, the samples for the fourth-grade assessments in each state consisted of about 2,500 fourth-graders in mathematics and about 2,500 in reading, from about 100 public schools in each case.

The target populations for the 1992 Trial State Assessment Program included only students in regular public schools[1] who were enrolled in the fourth or eighth grade at the time of assessment. The sampling frame included the public schools having the relevant grade (fourth or eighth grade) in each state or territory. The samples were selected based on a two-stage sample design—selection of schools within participating states and selection of students within schools. The first-stage samples of schools were selected with probability proportional to the eighth- or fourth-grade enrollment in the schools to provide efficient sample designs for the student populations. Special procedures were used for states with many small schools, and for states or territories having a small number of schools for a given grade (see section 3.4.5).

The sampling frame for each state was first stratified by the urbanization status of the area in which the school was located. The urbanization classes were defined in terms of large or mid-size central city, urban fringe of large or mid-size city, large town, small town, and rural areas (see section 3.4.2). Within urbanization strata, schools were further stratified explicitly on the basis of minority enrollment in those states with substantial Black or Hispanic student populations. Minority enrollment was defined as the total percent of Black and Hispanic students enrolled in a school (see section 3.4.3). Within minority strata, schools were sorted by median household income of the ZIP code area where the school was located (see section 3.4.4).

---

[1] A public school is defined as an institution which provides educational services and has one or more grade groups (PK-12) or which is ungraded, has one or more teachers to give instruction, is located in one or more buildings, has an assigned administrator, receives public funds as primary support, and is operated by an education agency. A regular school is a public elementary/secondary school that does not focus primarily on vocational, special, or alternative education.

One of the goals of the 1992 state sample design was to minimize overlap—between the state and national samples, between the state fourth- and eighth-grade samples (in schools that had both grades), and with the first phase followup to *Prospects: The National Longitudinal Study of Chapter I Children* (Abt Associates, 1991).

A systematic random sample of about 100 eighth-grade schools was drawn with probability proportional to the eighth-grade enrollment of the school from the stratified frame of schools within each state. Up to three sessions were assigned within each school. The number of sessions selected in each school was proportional to the eighth-grade enrollment of the schools. In those states and territories that had fewer than 100 schools with eighth grade, all schools were included in the sample.

Similarly, systematic random samples of fourth-grade schools were selected with probability proportional to the fourth-grade enrollment of the school from the fourth-grade sampling frames in the participating states. The number of schools drawn for the fourth-grade sample varied by state depending on the distribution of the fourth-grade enrollment in each state (see Table 3-3). In those states and territories that had fewer than 100 schools with fourth grade, all schools were included in the sample.

Successive schools were paired, using the same order in which they were selected, and one member of each pair was designated at random to be monitored during the assessment by Westat field staff so that reliable comparisons could be made between sessions administered with and without monitoring.

Both reading and mathematics sessions were conducted in fourth-grade sampled schools in which there were more than 20 students. Schools that had no more than 20 fourth-grade students were randomly assigned to administer either reading or mathematics. Approximately 2,500 students were assessed for each subject and each grade in a given state. Except in the two small territories, about 5,000 fourth-grade students participated in the assessment. On average, 128 fourth-grade schools were sampled in each state (in which sampling of schools was conducted) with about 115 conducting both mathematics and reading assessments, and about 13 conducting only mathematics or reading. The maximum number of schools selected in a state was 200.

Each selected school provided a list of eligible enrolled students, from which a systematic sample of students was drawn. Thirty students were selected for each session from grade 8 student lists, and 60 students were selected from grade 4 student lists. All students were selected if there were less than 30 grade 8 or less than 60 grade 4 students on the lists. Selected students within each of the fourth-grade schools were alternately assigned to either the mathematics or the reading assessments.

The 1992 assessment was preceded in 1991 by a field test, the principal goals of which were to test procedures and new items contemplated for the 1992 assessment. Three states and one territory also used the field test to observe and react to proposed strategies. Twenty-four jurisdictions participated in the field test. Schools that participated in the field test were given a chance of selection in the 1992 assessment, and there was no attempt to control the overlap between the school samples for the 1991 field test and those for the 1992 assessment. Section 3.2 documents the procedures used to select the schools for the field test.

Section 3.3 describes the construction of the sampling frames, including the sources of school data, missing data problems, and definition of in-scope schools. Section 3.4 includes a description of the various steps in stratification of schools within participating states. School sample selection procedures (including new and substitute schools) are described in section 3.5. Section 3.6 includes the steps involved in selection of students within participating schools.

3.2 SAMPLE SELECTION FOR THE 1991 FIELD TEST

The Trial State Assessment 1991 field test was conducted together with the field test for the national portion of the assessment. Twenty-four states participated in the field test, which was conducted for grades 4, 8, and 12. Pairs of schools were identified, with one of each pair to be included in the test. This allowed state participation in the selection of the test schools and also facilitated replacement of schools that declined to participate in the assessment. Sampling weights were not computed for the field test samples.

3.2.1 Primary Sampling Units

The frame of field-test PSUs was derived from the frame of NAEP PSUs[2], splitting PSUs where necessary in such a way that each of the new PSUs was completely contained within a single state. Each state was stratified by urbanization/minority. The sample sizes were assigned in such a way that for each NAEP region the sample sizes were proportional to the population of the participating states. Two PSUs were selected from each state. From each of the state strata, once the sample was assigned, the PSUs were selected with probability proportional to the 1980 population counts. The PSUs selected as noncertainties in the NAEP 1990 national sample were excluded from the PSU frame to avoid undue burden on the schools and districts in these PSUs. Controlled selection (Kish, 1965, pp. 488-95) of PSUs was used to achieve the selection of two PSUs per state, assigned proportionately among strata within each region.

Since two PSUs were selected for each of the participating states, the sample assignment was not proportional to the population counts. Overall, within each region, the assignment of PSUs was proportional to the urbanization/minority stratum population in each region, where the urbanization/minority stratum population distribution was based only upon the participating states, with each state contributing equally. So, for example, the rural population had disproportionately higher representation in the field test than in the general population, since many of the participating states were relatively rural in nature.

---

[2] The frame of NAEP PSUs was the frame used to draw the national NAEP samples for 1986 to 1992. Refer to the 1990 national technical report (Johnson & Allen, 1992) for more information.

3.2.2 Selection of Schools and Students

Public schools with fourth- and eighth-grade students were in scope for the assessment. Schools with fewer than 25 students per grade were eliminated from the frame, to eliminate the relatively high cost per student of conducting assessments in small schools.

The selection of schools avoided overlap with schools that had been selected from the certainty PSUs for the 1990 NAEP national sample and the IEA Reading Literacy Study, conducted for the National Center for Education Statistics (Rust & Bryant, 1991). Also, there was no overlap among the different grade samples.

For each grade, from each PSU, a sample of five schools was selected with probability proportional to the grade enrollment. In the states where one PSU had fewer than five schools, the sample from the other PSU in the state was increased so that the overall state sample was still 10 schools per grade. For each school, where the size of the PSU allowed, we selected the second member of each pair in such a way that the "distance" from the primary selection, based on percent of Black students, percent of Hispanic students, grade enrollment, and percent of students living below the poverty line, was the smallest. The overlap of samples was avoided by first selecting the twelfth-grade sample (for the national NAEP field test), then eliminating the selected schools from the eighth-grade sample selection, and then eliminating the twelfth- and eighth-grade selections before selecting the fourth-grade sample.

From each of the 10 schools selected for the eighth-grade sample, two classrooms were randomly selected from each of the five largest schools and one from each of the remaining schools. In the fourth-grade sample, two classrooms were selected randomly from each of the three largest schools and one classroom from each of the remaining seven schools. An exception was made in the fourth-grade samples in Florida, Kentucky, and Wisconsin, where 50 students were sampled from each of the three largest schools and 25 students from each of the remaining schools (unless the number of students was fewer than 35, in which case all of the them were taken in the sample). These three states wished to try out the student sampling procedures proposed for the 1992 assessment, and so did not use samples of intact classrooms.

3.2.3 Assignment to Sessions for Different Subjects

Three types of sessions were assigned for the field test: print-administered mathematics, audiotape-administered mathematics, and print-administered reading and writing. At grade 4, one classroom (session) per PSU was selected with equal probability to be administered the print-administered mathematics assessment in all states but Florida, Kentucky, and Wisconsin, for a total of 61 such sessions. The remaining 228 sessions were assigned to reading and writing, from which 15 sessions were selected for audiotaped mathematics sessions with equal probability after implicitly stratifying by geographic and urbanization/minority characteristics. In Florida, Kentucky, and Wisconsin, where samples of 50 students were drawn from the selected schools, the sample was randomly split in two equal sessions. Half of the sessions were randomly assigned to the print-administered mathematics assessment and the rest to the reading assessment. Florida, Kentucky, Wisconsin, and the Virgin Islands did not participate in any audiotaped mathematics sessions or writing sessions, since those two components were not planned to be part of the 1992 Trial State Assessment.

At grade 8, 50 sessions (classrooms)—two per state—were selected to conduct the print-administered mathematics assessment. The remaining 267 sessions were assigned to reading and writing, from which 13 sessions were assigned to tape-administered mathematics, subsampled with equal probability after implicitly stratifying by geographic and urbanization/minority characteristics. The Virgin Islands was the only jurisdiction at grade 8 that did not participate in the assignment of audiotaped sessions.

3.3 SAMPLING FRAME FOR THE 1992 ASSESSMENT

3.3.1 Choice of School Sampling Frame

In order to draw the school samples for the 1992 Trial State Assessment, it was necessary to obtain a comprehensive list of public schools in each state. For each school, we needed useful information for stratification purposes, reliable information about grade span and enrollment, and accurate information for identifying the school to the state coordinator (district membership, name, address).

Based on our experience with the 1990 Trial State Assessment, and national assessments in 1984, 1986, 1988, and 1990, we elected to use the file made available by Quality Education Data, Inc. (QED). We used the National Center for Education Statistics' Common Core of Data (CCD) school file to check the completeness of the QED file. This approach differed from that used to develop frames for the 1990 Trial State Assessment, for which the CCD was used primarily. There were several reasons for this change.

For 1992, it was possible to obtain a version of the QED file that contained all of the relevant variables from the most current CCD file. This meant in particular that data on minority enrollment by school, an important school stratification variable, were available on the QED file. These data had been available only on the CCD for the 1990 assessment. In addition, "type of locale," a seven-level urbanization variable newly created by the National Center for Education Statistics, was available on the QED (as well as the CCD) for 1992. Our experience in 1990 indicated that, generally speaking, the updatedness of the school lists and the quality of name and address information was both higher overall and more uniform on the QED. This is important for three reasons: 1) an outdated list leads to the selection of relatively many out-of-scope schools and greater reliance on new school sampling procedures; 2) poor quality name and address information leads to errors in the identification of sample schools by state coordinators (some schools on the CCD in 1990 had no city name as part of the address, for example); and 3) good quality ZIP codes are needed to give good stratification by household income (see section 3.4.4).

Thus, the combination of these factors led us to choose the QED file as the basis of the frame for each state. The QED list covers all states and territories except Puerto Rico (which did not participate). The version of the QED file used was released in late 1990, in time for selection of the school sample in early 1991. The file was missing minority and urbanization data for a sizable minority of schools (due to the inability of QED to match these schools with the corresponding CCD file). We undertook considerable efforts to obtain these variables for all schools in states where these variables were to be used for stratification. These efforts are described in the next section.

Tables 3-1 and 3-2 show the distribution of fourth- and eighth-grade schools, and enrollment within schools as reported in the 1990 QED file. Enrollment was estimated for each grade as the ratio of total school enrollment by the number of grades in the school. Refer to section 3.4.5 for the definition of small school cluster type. Schools with fewer than 20 students are denoted as small schools. Large schools are those with 20 or more students in the associated grade.

3.3.2 Missing Minority and Urbanization (Type of Locale) Data

As stated earlier, the sampling frame for the 1992 Trial State Assessment was the most recent version of the QED file, as of January 1991. The CCD file was used to extract information on minority and urbanization in the cases where these variables were missing on the QED file. The minority data were extracted only for those schools in states in which minority stratification was performed. In cases where urbanization could not be determined from the CCD file, the three-level classification of urban/suburban/rural (available for all schools on the QED file) was used to impute for urbanization.

3.3.3 In-scope Schools

The target population for the 1992 Trial State Assessment Program included students in regular public schools who were enrolled in the eighth grade or fourth grade. Parochial, private, Bureau of Indian Affairs, Department of Defense, and special education schools were not included.

3.4 WITHIN-STATE STRATIFICATION

3.4.1 Stratification Variables

Selection of schools within participating states involved three stages of explicit stratification and one stage of implicit stratification. The first three stages were school size (where size was the grade level enrollment of the schools), urbanization, and minority enrollment. The final stage was median income. The stratification methods described below applied to both fourth- and eighth-grade.

The first stage of stratification applied only to states with relatively many students in small schools. These states were known as Cluster Type 3 states. The schools were stratified into two strata, one stratum consisting of schools with 20 or more fourth-grade (or eighth-grade) students, and another stratum consisting of all schools with fewer than 20 students in the fourth (or eighth) grade. The primary purpose of this stratification was to ensure that the sample of schools would provide an appropriate student sample size. It also ensured appropriate

Table 3-1
Distribution of Fourth-grade Schools and Enrollment as Reported in QED 1990

State	Small School Cluster Type	Total Schools	Small Schools	Large Schools	Total Enrollment	Small School Enrollment
Alabama	Geographic	786	29	757	59,127	438
Arizona	Geographic	637	57	580	51,261	505
Arkansas	Geographic	550	40	510	35,107	606
California	Geographic	4,610	299	4,311	383,265	2,830
Colorado	Geographic	752	84	668	45,845	860
Connecticut	Geographic	563	11	552	37,069	163
Delaware	None	54	2	52	6,842	32
District of Columbia	None	118	3	115	6,206	34
Florida	Geographic	1,321	13	1,308	144,789	191
Georgia	Geographic	1,021	11	1,010	94,572	178
Guam	None	21	0	21	2,115	0
Hawaii	Geographic	170	3	167	14,070	21
Idaho	Stratified	304	44	258	18,069	385
Indiana	Geographic	1,167	21	1,146	75,807	339
Iowa	Stratified	794	84	710	37,786	1,236
Kentucky	Geographic	832	51	781	50,856	753
Louisiana	Geographic	788	44	744	62,780	627
Maine	Stratified	405	122	283	16,616	1,358
Maryland	Geographic	755	12	743	54,316	155
Massachusetts	Geographic	1,038	28	1,010	64,274	390
Michigan	Geographic	1,876	62	1,814	123,028	571
Minnesota	Geographic	838	66	772	58,711	956
Mississippi	Geographic	465	3	462	41,063	46
Missouri	Stratified	1,093	147	946	63,555	1,728
Nebraska	Stratified	1,011	615	396	21,834	3,226
New Hampshire	Stratified	268	55	213	13,721	654
New Jersey	Geographic	1,338	42	1,296	84,148	639
New Mexico	Stratified	378	57	321	24,316	673
New York	Geographic	2,259	44	2,215	191,873	565
North Carolina	Geographic	1,109	25	1,084	85,158	361
North Dakota	Stratified	359	180	179	9,973	1,628
Ohio	Geographic	2,039	44	1,995	136,626	651
Oklahoma	Stratified	973	216	757	48,217	2,696
Pennsylvania	Geographic	1,879	47	1,832	126,166	727
Rhode Island	Geographic	177	2	175	11,114	28
South Carolina	Geographic	552	4	548	49,117	50
Tennessee	Geographic	933	66	867	66,932	900
Texas	Geographic	3,053	238	2,815	268,796	2,896
Utah	Geographic	432	31	401	36,629	260
Virginia	Geographic	1,041	39	1,002	80,886	523
Virgin Islands	None	24	1	23	1,874	15
West Virginia	Stratified	637	104	533	25,532	1,474
Wisconsin	Stratified	1,147	128	1,019	59,965	1,910
Wyoming	Stratified	238	91	147	8,050	528

Table 3-2
Distribution of Eighth-grade Schools and Enrollment as Reported in QED 1990

State	Small School Cluster Type	Total Schools	Small Schools	Large Schools	Total Enrollment	Small School Enrollment
Alabama	Geographic	497	15	482	55,735	231
Arizona	Geographic	303	42	261	44,533	349
Arkansas	Geographic	358	32	326	34,237	461
California	Geographic	1,594	214	1,380	330,433	2,023
Colorado	Geographic	319	62	257	40,763	637
Connecticut	Geographic	214	6	208	31,483	63
Delaware	None	28	2	26	6,482	32
District of Columbia	None	35	0	35	5,361	0
Florida	Geographic	442	7	435	125,900	84
Georgia	Geographic	393	1	392	86,778	10
Guam	None	6	0	6	1,862	0
Hawaii	None	57	2	55	12,053	21
Idaho	Geographic	154	29	125	16,243	265
Indiana	Geographic	444	2	442	72,226	27
Iowa	Geographic	455	37	418	36,272	467
Kentucky	Geographic	432	34	398	47,605	501
Louisiana	Geographic	440	45	395	57,168	648
Maine	Stratified	235	68	167	15,713	713
Maryland	Geographic	216	4	212	48,408	41
Massachusetts	Geographic	385	4	381	58,519	22
Michigan	Geographic	748	42	706	113,633	331
Minnesota	Geographic	441	30	411	53,079	444
Mississippi	Geographic	308	2	306	37,965	31
Missouri	Stratified	640	131	509	58,673	1,519
Nebraska	Stratified	706	502	204	19,986	2,636
New Hampshire	Geographic	134	16	118	12,787	184
New Jersey	Geographic	678	23	655	81,852	332
New Mexico	Geographic	150	28	122	21,111	310
New York	Geographic	998	15	983	185,484	196
North Carolina	Geographic	556	17	539	84,003	225
North Dakota	Stratified	272	167	105	8,809	1,555
Ohio	Geographic	846	16	830	129,321	180
Oklahoma	Stratified	642	207	435	44,121	2,559
Pennsylvania	Geographic	722	2	720	122,456	19
Rhode Island	None	54	1	53	9,765	9
South Carolina	Geographic	256	1	255	47,670	12
Tennessee	Geographic	549	47	502	63,532	611
Texas	Geographic	1,464	210	1,254	242,858	2,567
Utah	Geographic	142	17	125	31,840	162
Virginia	Geographic	335	5	330	72,407	67
Virgin Islands	None	6	0	6	1,960	0
West Virginia	Geographic	252	14	238	25,206	204
Wisconsin	Geographic	517	42	475	54,906	605
Wyoming	Stratified	108	47	61	7,358	267

representation of small schools in states with any substantial number of such schools. Tables 3-3 and 3-4 provide the type of stratification used in each of the participating states or territories respectively for fourth- and eighth-grade samples. Refer to section 3.4.2 for the definition of urbanization and section 3.4.3 for the definition of minority.

3.4.2 Urbanization Classification

The NCES "type of locale" variable was used to stratify schools into different urbanization levels. Stratification by type of locale was repeated separately for fourth and eighth grade. The seven categories of the variable are defined as follows.

1) *Large Central City:* a central city of a Metropolitan Statistical Area (MSA) with a population greater than or equal to 400,000, or a population density greater than or equal to 6,000 persons per square mile.

2) *Mid-size Central City:* a central city of an MSA but not designated as a large central city.

3) *Urban Fringe of Large City:* a place within an MSA of a large central city and defined as urban by the U.S. Bureau of Census.

4) *Urban Fringe of Mid-size City:* a place within an MSA of a mid-size central city and defined as urban by the U.S. Bureau of Census.

5) *Large Town:* a place not within an MSA, but with a population greater than or equal to 25,000 and defined as urban by the U.S. Bureau of Census.

6) *Small Town:* a place not within an MSA, but with a population less than 25,000 and defined as urban by U.S. Bureau of Census.

7) *Rural:* a place with a population of less than 2,500 and defined as rural by the U.S. Bureau of the Census.

The urbanization strata were created by collapsing type of locale categories. The nature of the collapsing varied across states and grades. At a minimum, each urbanization stratum included 10 percent of eligible students in the participating state. Tables 3-3 and 3-4 provide the urbanization categories (created by collapsing type of locale) used within each state.

3.4.3 Minority Classification

The third stage of stratification was minority enrollment. Minority enrollment strata were formed within urbanization strata, based on the percentages of Black and Hispanic students. The three cases that occur are described in the following paragraphs.

Case 1: Urbanization strata with less than 10 percent Black students and 7 percent Hispanic students were not stratified by minority enrollment.

Table 3-3
Distribution of the Selected Schools by Sampling Strata, Grade 4

Urbanization	Minority	Schools in Strata
ALABAMA (Small School Cluster Type 2 - Geographic)		
Mid-size Central City	Low Percent Minority	9
Mid-size Central City	Medium Percent Minority	9
Mid-size Central City	High Percent Minority	8
Urban Fringe of Mid-size Central City	Low Percent Minority	10
Urban Fringe of Mid-size Central City	Medium Percent Minority	10
Urban Fringe of Mid-size Central City	High Percent Minority	9
Large/Small Town	High Percent Minority	9
Large/Small Town	Low Percent Minority	9
Large/Small Town	Medium Percent Minority	9
Rural	Low Percent Minority	16
Rural	Medium Percent Minority	14
		112
ARIZONA (Small School Cluster Type 2 - Geographic)		
Large Central City	Low Percent Minority	9
Large Central City	Medium Percent Minority	8
Large Central City	High Percent Minority	9
Mid-size Central City	Low Percent Minority	10
Mid-size Central City	Medium Percent Minority	10
Mid-size Central City	High Percent Minority	9
Urban Fringe of Large Central City	Low Percent Minority	6
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	7
Urban Fringe of Large/Mid-size Central City	High Percent Minority	6
Large/Small Town and Rural	Low Percent Minority	13
Large/Small Town and Rural	Medium Percent Minority	13
Large/Small Town and Rural	High Percent Minority	10
		110
ARKANSAS (Small School Cluster Type 2 - Geographic)		
Mid-size Central City+Urban Fringe	Low Percent Minority	10
Mid-size Central City+Urban Fringe	Medium Percent Minority	10
Mid-size Central City+Urban Fringe	High Percent Minority	10
L/Small Town	Low Percent Minority	15
L/Small Town	Medium Percent Minority	14
L/Small Town	High Percent Minority	15
Rural	Low Percent Minority	19
Rural	Medium Percent Minority	15
Rural	High Percent Minority	16
		124

Table 3-3 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 4

Urbanization	Minority	Schools in Strata
CALIFORNIA (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	13
Large/Mid-size Central City	Medium Percent Minority	12
Large/Mid-size Central City	High Percent Minority	13
Urban Fringe of Large Central City	Low Percent Minority	10
Urban Fringe of Large Central City	Medium Percent Minority	11
Urban Fringe of Large Central City	High Percent Minority	11
Urban Fringe of Mid-size Central City	Low Percent Minority	5
Urban Fringe of Mid-size Central City	Medium Percent Minority	5
Urban Fringe of Mid-size Central City	High Percent Minority	4
Large/Small Town and Rural	Low Percent Minority	13
Large/Small Town and Rural	Medium Percent Minority	9
Large/Small Town and Rural	High Percent Minority	7
		113
COLORADO (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	12
Large/Mid-size Central City	Medium Percent Minority	11
Large/Mid-size Central City	High Percent Minority	11
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	14
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	14
Urban Fringe of Large/Mid-size Central City	High Percent Minority	14
Large/Small Town	Low Percent Minority	7
Large/Small Town	Medium Percent Minority	7
Large/Small Town	High Percent Minority	6
Rural	Low Percent Minority	11
Rural	Medium Percent Minority	9
Rural	High Percent Minority	11
		127
CONNECTICUT (Small School Cluster Type 2 - Geographic)		
Large Central City	Low Black/Low Hispanic	5
Large Central City	Low Black/High Hispanic	4
Large Central City	High Black/Low Hispanic	4
Large Central City	High Black/High Hispanic	4
Mid-size Central City	Low Percent Minority	7
Mid-size Central City	Medium Percent Minority	7
Mid-size Central City	High Percent Minority	7
Urban Fringe of Large/Mid-size Central City	None	34
Large/Small Town and Rural	None	39
		111

Table 3-3 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 4

Urbanization	Minority	Schools in Strata
DELAWARE (Small School Cluster Type 1 - None)		
Large/Mid-size Central City	Low Percent Minority	3
Large/Mid-size Central City	Medium Percent Minority	4
Large/Mid-size Central City	High Percent Minority	5
Urban Fringe of Mid-size Central City	Low Percent Minority	1
Urban Fringe of Mid-size Central City	Medium Percent Minority	2
Urban Fringe of Mid-size Central City	High Percent Minority	3
Small Town	Low Percent Minority	2
Small Town	Medium Percent Minority	3
Small Town	High Percent Minority	1
Rural	Low Percent Minority	7
Rural	Medium Percent Minority	8
Rural	High Percent Minority	8
		47
DISTRICT OF COLUMBIA (Small School Cluster Type 1 - None)		
Large Central City	Medium Percent Minority	44
Large Central City	High Percent Minority	69
		113
FLORIDA (Small School Cluster Type 2 - Geographic)		
Large Central City	Low Black/Low Hispanic	4
Large Central City	Low Black/High Hispanic	4
Large Central City	High Black/Low Hispanic	4
Large Central City	High Black/High Hispanic	4
Mid-size Central City	Low Percent Minority	6
Mid-size Central City	Medium Percent Minority	7
Mid-size Central City	High Percent Minority	7
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	16
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	16
Urban Fringe of Large/Mid-size Central City	High Percent Minority	15
Large/Small Town and Rural	Low Percent Minority	8
Large/Small Town and Rural	Medium Percent Minority	8
Large/Small Town and Rural	High Percent Minority	7
		106
GEORGIA (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	8
Large/Mid-size Central City	Medium Percent Minority	8
Large/Mid-size Central City	High Percent Minority	8
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	10
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	11
Urban Fringe of Large/Mid-size Central City	High Percent Minority	10
Large/Small Town	Low Percent Minority	11
Large/Small Town	Medium Percent Minority	11
Large/Small Town	High Percent Minority	10
Rural	Low Percent Minority	7
Rural	Medium Percent Minority	6
Rural	High Percent Minority	7
		107

Table 3-3 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 4

Urbanization	Minority	Schools in Strata
GUAM (Small School Cluster Type 1 - None)		
Rural	None	21
HAWAII (Small School Cluster Type 2 - Geographic)		
Mid-size Central City	None	33
Urban Fringe of Mid-size Central City	None	51
Large/Small Town and Rural	None	23
		107
IDAHO (Small School Cluster Type 3 - Stratified)		
Large Schools		
Mid-size Central City and Urban Fringe	None	22
Large Town	None	19
Small Town	None	35
Rural	None	39
Small Schools		
None	None	14
		129
INDIANA (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	12
Large/Mid-size Central City	Medium Percent Minority	11
Large/Mid-size Central City	High Percent Minority	10
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	13
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	11
Rural	None	26
Large/Small Town	None	33
		116
IOWA (Small School Cluster Type 3 - Stratified)		
Large Schools		
Mid-size Central City and Urban Fringe	None	38
Large/Small Town	None	40
Rural	None	47
Small Schools		
None	None	14
		139
KENTUCKY (Small School Cluster Type 2 - Geographic)		
Mid-size Central City	Low Percent Minority	6
Mid-size Central City	Medium Percent Minority	7
Mid-size Central City	High Percent Minority	6
Urban Fringe of Mid-size Central City	Low Percent Minority	9
Urban Fringe of Mid-size Central City	Medium Percent Minority	8
Rural	None	51
Large/Small Town	None	36
		123

Table 3-3 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 4

Urbanization	Minority	Schools in Strata
LOUISIANA (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	11
Large/Mid-size Central City	Medium Percent Minority	11
Large/Mid-size Central City	High Percent Minority	12
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	6
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	7
Urban Fringe of Large/Mid-size Central City	High Percent Minority	6
Large/Small Town	Low Percent Minority	11
Large/Small Town	Medium Percent Minority	11
Large/Small Town	High Percent Minority	11
Rural	Low Percent Minority	8
Rural	Medium Percent Minority	11
Rural	High Percent Minority	9
		114
MAINE (Small School Cluster Type 3 - Stratified)		
Large Schools		
Mid-size Central City and Urban Fringe	None	21
Small Town	None	58
Rural	None	45
Small Schools		
None	None	39
		163
MARYLAND (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	7
Large/Mid-size Central City	Medium Percent Minority	6
Large/Mid-size Central City	High Percent Minority	7
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	21
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	22
Urban Fringe of Large/Mid-size Central City	High Percent Minority	21
Large/Small Town and Rural	Low Percent Minority	14
Large/Small Town and Rural	Medium Percent Minority	12
		110
MASSACHUSETTS (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	14
Large/Mid-size Central City	Medium Percent Minority	13
Large/Mid-size Central City	High Percent Minority	13
Urban Fringe of Large/Mid-size Central City	None	40
Large/Small Town and Rural	None	40
		120
MICHIGAN (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	9
Large/Mid-size Central City	Medium Percent Minority	8
Large/Mid-size Central City	High Percent Minority	9
Urban Fringe of Large/Mid-size Central City	None	38
Rural	None	20
Large/Small Town	None	30
		114

Table 3-3 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 4

Urbanization	Minority	Schools in Strata
MINNESOTA (Small School Cluster Type 2 - Geographic)		
Large Central City	Medium Percent Minority	5
Mid-size Central City	Low Percent Minority	7
Urban Fringe of Large/Mid-size Central City	None	36
Rural	None	41
Large/Small Town	None	26
		115
MISSISSIPPI (Small School Cluster Type 2 - Geographic)		
Mid-size Central City	Low Percent Minority	4
Mid-size Central City	Medium Percent Minority	5
Mid-size Central City	High Percent Minority	4
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	3
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	3
Urban Fringe of Large/Mid-size Central City	High Percent Minority	3
Large/Small Town	Low Percent Minority	15
Large/Small Town	Medium Percent Minority	15
Large/Small Town	High Percent Minority	15
Rural	Low Percent Minority	13
Rural	Medium Percent Minority	13
Rural	High Percent Minority	15
		108
MISSOURI (Small School Cluster Type 3 - Stratified)		
Large Schools		
Large/Mid-size Central City	Low Percent Minority	5
Large/Mid-size Central City	Medium Percent Minority	6
Large/Mid-size Central City	High Percent Minority	3
Urban Fringe of Large Central City	High Percent Minority	13
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	13
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	13
Large/Small Town	None	24
Rural	None	33
Small Schools		
None	None	13
		123
NEBRASKA (Small School Cluster Type 3 - Stratified)		
Large Schools		
Mid-size Central City and Urban Fringe	None	43
Large/Small Town	None	37
Rural	None	41
Small Schools		
None	None	66
		187

Table 3-3 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 4

Urbanization	Minority	Schools in Strata
NEW HAMPSHIRE (Small School Cluster Type 3 - Stratified)		
Large Schools Mid-size Central City and Urban Fringe	None	26
Large/Small Town	None	57
Rural	None	29
Small Schools None	None	24
		136
NEW JERSEY (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Black/Low Hispanic	6
Large/Mid-size Central City	Low Black/High Hispanic	5
Large/Mid-size Central City	High Black/Low Hispanic	5
Large/Mid-size Central City	High Black/High Hispanic	5
Urban Fringe of Large Central City	Low Percent Minority	28
Urban Fringe of Large Central City	Medium Percent Minority	17
Urban Fringe of Mid-size Central City	None	25
Large/Small Town and Rural	None	28
		119
NEW MEXICO (Small School Cluster Type 3 - Stratified)		
Large Schools		
Mid-size Central City and Urban Fringe	Low Percent Minority	14
Mid-size Central City and Urban Fringe	Medium Percent Minority	14
Mid-size Central City and Urban Fringe	High Percent Minority	14
Large Town	Low Percent Minority	5
Large Town	Medium Percent Minority	5
Large Town	High Percent Minority	6
Small Town	Low Percent Minority	10
Small Town	Medium Percent Minority	10
Small Town	High Percent Minority	11
Rural	Low Percent Minority	5
Rural	Medium Percent Minority	7
Rural	High Percent Minority	8
Small Schools		
None	None	11
		120
NEW YORK (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	High Black/High Hispanic	11
Large/Mid-size Central City	Low Black/Low Hispanic	12
Large/Mid-size Central City	Low Black/High Hispanic	12
Large/Mid-size Central City	High Black/Low Hispanic	12
Urban Fringe of Large Central City	None	13
Urban Fringe of Mid-size Central City	None	18
Large/Small Town and Rural	None	32
		110

Table 3-3 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 4

Urbanization	Minority	Schools in Strata
NORTH CAROLINA (Small School Cluster Type 2 - Geographic)		
Mid-size Central City	Low Percent Minority	10
Mid-size Central City	Medium Percent Minority	9
Mid-size Central City	High Percent Minority	10
Urban Fringe of Mid-size Central City	Low Percent Minority	4
Urban Fringe of Mid-size Central City	Medium Percent Minority	5
Urban Fringe of Mid-size Central City	High Percent Minority	4
Large/Small Town	Low Percent Minority	11
Large/Small Town	Medium Percent Minority	11
Large/Small Town	High Percent Minority	10
Rural	Low Percent Minority	17
Rural	Medium Percent Minority	12
Rural	High Percent Minority	12
		115
NORTH DAKOTA (Small School Cluster Type 3 - Stratified)		
Large Schools		
Mid-size Central City and Urban Fringe	None	36
Large/Small Town	None	31
Rural	None	51
Small Schools		
None	None	42
		160
OHIO (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	11
Large/Mid-size Central City	Medium Percent Minority	10
Large/Mid-size Central City	High Percent Minority	11
Urban Fringe of Large/Mid-size Central City	None	32
Large/Small Town	None	24
Rural	None	29
		117
OKLAHOMA (Small School Cluster Type 3 - Stratified)		
Large Schools		
Large/Mid-size Central City	Low Percent Minority	16
Large/Mid-size Central City	Medium Percent Minority	17
Urban Fringe of Large/Mid-size Central City	None	14
Large/Small Town	None	37
Rural	None	34
Small Schools None	None	23
		141

Table 3-3 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 4

Urbanization	Minority	Schools in Strata
PENNSYLVANIA (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	9
Large/Mid-size Central City	Medium Percent Minority	9
Large/Mid-size Central City	High Percent Minority	8
Urban Fringe of Large/Mid-size Central City	None	33
Large/Small Town	None	36
Rural	None	23
		118
RHODE ISLAND (Small School Cluster Type 2 - Geographic)		
Large Central City	Low Percent Minority	8
Large Central City	Medium Percent Minority	6
Large Central City	High Percent Minority	5
Mid-size Central City	None	9
Urban Fringe of Large/Mid-size Central City	None	55
Large/Small Town and Rural	None	27
		110
SOUTH CAROLINA (Small School Cluster Type 2 - Geographic)		
Mid-size Central City	Low Percent Minority	6
Mid-size Central City	Medium Percent Minority	5
Mid-size Central City	High Percent Minority	6
Urban Fringe of Mid-size Central City	Low Percent Minority	10
Urban Fringe of Mid-size Central City	Medium Percent Minority	10
Urban Fringe of Mid-size Central City	High Percent Minority	10
Small Town	Low Percent Minority	13
Small Town	Medium Percent Minority	12
Small Town	High Percent Minority	12
Rural	Low Percent Minority	9
Rural	Medium Percent Minority	9
Rural	High Percent Minority	9
		111
TENNESSEE (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	13
Large/Mid-size Central City	Medium Percent Minority	13
Large/Mid-size Central City	High Percent Minority	12
Urban Fringe of Large/Mid-size Central City	None	19
Large/Small Town	None	31
Rural	None	32
		120

Table 3-3 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 4

Urbanization	Minority	Schools in Strata
TEXAS (Small School Cluster Type 2 - Geographic)		
Large Central City	Low Hispanic/Low Black	7
Large Central City	Low Hispanic/High Black	6
Large Central City	High Hispanic/Low Black	7
Large Central City	High Hispanic/High Black	6
Mid-size Central City	Low Percent Minority	7
Mid-size Central City	Medium Percent Minority	8
Mid-size Central City	High Percent Minority	9
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	7
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	5
Urban Fringe of Large/Mid-size Central City	High Percent Minority	5
Large/Small Town and Rural	Low Percent Minority	15
Large/Small Town and Rural	Medium Percent Minority	14
Large/Small Town and Rural	High Percent Minority	15
		111
UTAH (Small School Cluster Type 2 - Geographic)		
Mid-size Central City	None	26
Urban Fringe of Mid-size Central City	None	46
Large/Small Town	None	15
Rural	None	24
		111
VIRGINIA (Small School Cluster Type 2 - Geographic)		
Mid-size Central City	Low Percent Minority	13
Mid-size Central City	Medium Percent Minority	11
Mid-size Central City	High Percent Minority	12
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	11
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	10
Urban Fringe of Large/Mid-size Central City	High Percent Minority	9
Small Town	Medium Percent Minority	5
Small Town	High Percent Minority	6
Large/Small Town	Low Percent Minority	5
Rural	Low Percent Minority	9
Rural	Medium Percent Minority	11
Rural	High Percent Minority	8
		110
VIRGIN ISLANDS (Small School Cluster Type 1 - None)		
	Low Percent Minority	10
	Medium Percent Minority	6
	High Percent Minority	8
		24

Table 3-3 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 4

Urbanization	Minority	Schools in Strata
WEST VIRGINIA (Small School Cluster Type 3 - Stratified)		
Large Schools		
Mid-size Central City	None	18
Urban Fringe of Mid-size Central City	None	19
Large/Small Town	None	35
Rural	None	65
Small Schools		
None	None	19
		156
WISCONSIN (Small School Cluster Type 3 - Stratified)		
Large Schools		
Large/Mid-size Central City	Low Percent Minority	17
Large/Mid-size Central City	Medium Percent Minority	17
Urban Fringe of Large/Mid-size Central City	None	20
Large/Small Town	None	31
Rural	None	32
Small Schools		
None	None	12
		129
WYOMING (Small School Cluster Type 3 - Stratified)		
Large Schools		
Mid-size Central City	None	18
Urban Fringe of Mid-size Central City	None	15
Large/Small Town	None	62
Rural	None	25
Small Schools		
None	None	60
		180

Table 3-4
Distribution of the Selected Schools by Sampling Strata, Grade 8

Urbanization	Minority	Schools in Strata
ALABAMA (Small School Cluster Type 2 - Geographic)		
Mid-size Central City	Low Percent Minority	7
Mid-size Central City	Medium Percent Minority	6
Mid-size Central City	High Percent Minority	8
Urban Fringe of Mid-size Central City	Low Percent Minority	10
Urban Fringe of Mid-size Central City	Medium Percent Minority	9
Urban Fringe of Mid-size Central City	High Percent Minority	8
Small Town	Low Percent Minority	10
Small Town	Medium Percent Minority	9
Large/Small Town	High Percent Minority	9
Rural	Low Percent Minority	13
Rural	Medium Percent Minority	8
Rural	High Percent Minority	9
		106
ARIZONA (Small School Cluster Type 2 - Geographic)		
Large Central City	Low Percent Minority	8
Large Central City	Medium Percent Minority	8
Large Central City	High Percent Minority	8
Mid-size Central City	Low Percent Minority	9
Mid-size Central City	Medium Percent Minority	10
Mid-size Central City	High Percent Minority	9
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	5
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	6
Urban Fringe of Large/Mid-size Central City	High Percent Minority	5
Large/Small Town and Rural	Low Percent Minority	13
Large/Small Town and Rural	Medium Percent Minority	10
Large/Small Town and Rural	High Percent Minority	12
		103
ARKANSAS (Small School Cluster Type 2 - Geographic)		
Mid-size Central City and Urban Fringe	Low Percent Minority	9
Mid-size Central City and Urban Fringe	Medium Percent Minority	7
Mid-size Central City and Urban Fringe	High Percent Minority	8
Large/Small Town	Low Percent Minority	13
Large/Small Town	Medium Percent Minority	15
Large/Small Town	High Percent Minority	14
Rural	Low Percent Minority	14
Rural	Medium Percent Minority	9
Rural	High Percent Minority	11
		100

Table 3-4 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 8

Urbanization	Minority	Schools in Strata
CALIFORNIA (Small School Cluster Type 2 - Geographic)		
Large Central City	Low Percent Minority	7
Large Central City	Medium Percent Minority	7
Large Central City	High Percent Minority	8
Mid-size Central City	Low Percent Minority	5
Mid-size Central City	Medium Percent Minority	5
Mid-size Central City	High Percent Minority	5
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	16
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	15
Urban Fringe of Large/Mid-size Central City	High Percent Minority	16
Large/Small Town and Rural	Low Percent Minority	6
Large/Small Town and Rural	Medium Percent Minority	7
Large/Small Town and Rural	High Percent Minority	6
		103
COLORADO (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	11
Large/Mid-size Central City	Medium Percent Minority	12
Large/Mid-size Central City	High Percent Minority	11
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	12
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	12
Urban Fringe of Large/Mid-size Central City	High Percent Minority	13
Large/Small Town	Low Percent Minority	7
Large/Small Town	Medium Percent Minority	7
Large/Small Town	High Percent Minority	7
Rural	Low Percent Minority	7
Rural	Medium Percent Minority	6
Rural	High Percent Minority	7
		112
CONNECTICUT (Small School Cluster Type 2 - Geographic)		
Large Central City	Low Black/Low Hispanic	3
Large Central City	Low Black/High Hispanic	1
Large Central City	High Black/Low Hispanic	2
Large Central City	High Black/High Hispanic	4
Mid-size Central City	Low Percent Minority	5
Mid-size Central City	Medium Percent Minority	6
Mid-size Central City	High Percent Minority	7
Urban Fringe of Large/Mid-size Central City	None	30
Large/Small Town and Rural	None	40
		98

Table 3-4 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 8

Urbanization	Minority	Schools in Strata
DELAWARE (Small School Cluster Type 1 - None)		
Mid-size Central City	Low Percent Minority	1
Mid-size Central City	High Percent Minority	1
Urban Fringe of Mid-size Central City	Low Percent Minority	2
Urban Fringe of Mid-size Central City	Medium Percent Minority	4
Urban Fringe of Mid-size Central City	High Percent Minority	3
Small Town	Low Percent Minority	2
Small Town	Medium Percent Minority	2
Small Town	High Percent Minority	1
Rural	Low Percent Minority	3
Rural	Medium Percent Minority	3
Rural	High Percent Minority	4
		26
DISTRICT OF COLUMBIA (Small School Cluster Type 1 - None)		
Large Central City	Medium Percent Minority	14
Large Central City	High Percent Minority	19
		33
FLORIDA (Small School Cluster Type 2 - Geographic)		
Large Central City	Low Black Low Hispanic	2
Large Central City	Low Black High Hispanic	5
Large Central City	High Black Low Hispanic	3
Large Central City	High Black High Hispanic	4
Mid-size Central City	Low Percent Minority	7
Mid-size Central City	Medium Percent Minority	7
Mid-size Central City	High Percent Minority	6
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	16
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	15
Urban Fringe of Large/Mid-size Central City	High Percent Minority	14
Large/Small Town and Rural	Low Percent Minority	7
Large/Small Town and Rural	Medium Percent Minority	8
Large/Small Town and Rural	High Percent Minority	7
		101
GEORGIA (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	6
Large/Mid-size Central City	Medium Percent Minority	7
Large/Mid-size Central City	High Percent Minority	7
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	11
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	10
Urban Fringe of Large/Mid-size Central City	High Percent Minority	10
Large/Small Town	Low Percent Minority	10
Large/Small Town	Medium Percent Minority	12
Large/Small Town	High Percent Minority	11
Rural	Low Percent Minority	5
Rural	Medium Percent Minority	5
Rural	High Percent Minority	6
		100

Table 3-4 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 8

Urbanization	Minority	Schools in Strata
GUAM (Small School Cluster Type 1 - None)		
Rural	None	6
HAWAII (Small School Cluster Type 1 - None)		
Mid-size Central City	None	12
Urban Fringe of Mid-size Central City	None	22
Large/Small Town and Rural	None	18
		52
IDAHO (Small School Cluster Type 2 - Geographic)		
Mid-size Central City and Urban Fringe	None	10
Large Town	None	10
Small Town	None	26
Rural	None	35
		81
INDIANA (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	9
Large/Mid-size Central City	Medium Percent Minority	10
Large/Mid-size Central City	High Percent Minority	8
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	10
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	7
Urban Fringe of Large/Mid-size Central City	High Percent Minority	7
Large/Small Town	None	34
Rural	None	20
		105
IOWA (Small School Cluster Type 2 - Geographic)		
Mid-size Central City and Urban Fringe	None	32
Large/Small Town	None	38
Rural	None	35
		105
KENTUCKY (Small School Cluster Type 2 - Geographic)		
Mid-size Central City	Low Percent Minority	6
Mid-size Central City	Medium Percent Minority	6
Mid-size Central City	High Percent Minority	5
Urban Fringe of Mid-size Central City	Low Percent Minority	6
Urban Fringe of Mid-size Central City	Medium Percent Minority	4
Urban Fringe of Mid-size Central City	High Percent Minority	5
Rural	None	31
Large/Small Town	None	41
		104

Table 3-4 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 8

Urbanization	Minority	Schools in Strata
LOUISIANA (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	9
Large/Mid-size Central City	Medium Percent Minority	10
Large/Mid-size Central City	High Percent Minority	11
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	7
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	7
Urban Fringe of Large/Mid-size Central City	High Percent Minority	8
Large/Small Town	Low Percent Minority	10
Large/Small Town	Medium Percent Minority	11
Large/Small Town	High Percent Minority	10
Rural	Low Percent Minority	6
Rural	Medium Percent Minority	6
Rural	High Percent Minority	6
		101
MAINE (Small School Cluster Type 3 - Stratified)		
Large Schools		
Mid-size Central City and Urban Fringe	None	15
Small Town	None	50
Rural	None	25
Small Schools		
None	None	10
		100
MARYLAND (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	6
Large/Mid-size Central City	Medium Percent Minority	7
Large/Mid-size Central City	High Percent Minority	6
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	20
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	21
Urban Fringe of Large/Mid-size Central City	High Percent Minority	18
Large/Small Town and Rural	Low Percent Minority	10
Large/Small Town and Rural	Medium Percent Minority	8
Large/Small Town and Rural	High Percent Minority	7
		103
MASSACHUSETTS (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	11
Large/Mid-size Central City	Medium Percent Minority	9
Large/Mid-size Central City	High Percent Minority	10
Urban Fringe of Large/Mid-size Central City	None	29
Large/Small Town	None	39
		98

Table 3-4 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 8

Urbanization	Minority	Schools in Strata
MICHIGAN (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	7
Large/Mid-size Central City	Medium Percent Minority	8
Large/Mid-size Central City	High Percent Minority	8
Urban Fringe of Large/Mid-size Central City	None	36
Large/Small Town	None	30
Rural	None	16
		105
MINNESOTA (Small School Cluster Type 2 - Geographic)		
Large Central City	Medium Percent Minority	6
Mid-size Central City	Low Percent Minority	7
Urban Fringe of Large/Mid-size Central City	None	32
Large/Small Town	None	27
Rural	None	29
		101
MISSISSIPPI (Small School Cluster Type 2 - Geographic)		
Mid-size Central City	Low Percent Minority	4
Mid-size Central City	Medium Percent Minority	3
Mid-size Central City	High Percent Minority	4
Urban Fringe of Mid-size Central City	Low Percent Minority	3
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	3
Urban Fringe of Large/Mid-size Central City	High Percent Minority	4
Large/Small Town	Low Percent Minority	16
Large/Small Town	Medium Percent Minority	15
Large/Small Town	High Percent Minority	14
Rural	Low Percent Minority	11
Rural	Medium Percent Minority	12
Rural	High Percent Minority	10
		99
MISSOURI (Small School Cluster Type 3 - Stratified)		
Large Schools		
Large/Mid-size Central City	Low Percent Minority	4
Large/Mid-size Central City	Medium Percent Minority	4
Large/Mid-size Central City	High Percent Minority	4
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	11
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	12
Urban Fringe of Large/Mid-size Central City	High Percent Minority	11
Large/Small Town	None	28
Rural	None	26
Small Schools		
None	None	6
		106

Table 3-4 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 8

Urbanization	Minority	Schools in Strata
NEBRASKA (Small School Cluster Type 3 - Stratified)		
Large Schools		
Mid-size Central City and Urban Fringe	None	24
Large/Small Town	None	25
Rural	None	30
Small Schools		
None	None	32
		111
NEW HAMPSHIRE (Small School Cluster Type 2 - Geographic)		
Mid-size Central City and Urban Fringe	None	13
Large/Small Town	None	45
Rural	None	19
		77
NEW JERSEY (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Black/Low Hispanic	4
Large/Mid-size Central City	Low Black/High Hispanic	5
Large/Mid-size Central City	High Black/Low Hispanic	5
Large/Mid-size Central City	High Black/High Hispanic	5
Urban Fringe of Large Central City	Low Percent Minority	24
Urban Fringe of Large Central City	Medium Percent Minority	14
Urban Fringe of Mid-size Central City	None	24
Large/Small Town and Rural	None	24
		105
NEW MEXICO (Small School Cluster Type 2 - Geographic)		
Mid-size Central City and Urban Fringe	Low Percent Minority	10
Mid-size Central City and Urban Fringe	Medium Percent Minority	9
Mid-size Central City and Urban Fringe	High Percent Minority	10
Large Town	Low Percent Minority	4
Large Town	Medium Percent Minority	5
Large Town	High Percent Minority	5
Small Town	Low Percent Minority	9
Small Town	Medium Percent Minority	11
Small Town	High Percent Minority	9
Rural	Low Percent Minority	5
Rural	Medium Percent Minority	9
Rural	High Percent Minority	6
		92

Table 3-4 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 8

Urbanization	Minority	Schools in Strata
NEW YORK (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Black/High Hispanic	10
Large/Mid-size Central City	Low Black/Low Hispanic	11
Large/Mid-size Central City	High Black/Low Hispanic	10
Large/Mid-size Central City	High Black/High Hispanic	11
Urban Fringe of Large Central City	Low Percent Minority	8
Urban Fringe of Large Central City	Medium Percent Minority	7
Urban Fringe of Mid-size Central City	None	18
Large/Small Town and Rural	None	29
		104
NORTH CAROLINA (Small School Cluster Type 2 - Geographic)		
Mid-size Central City	Low Percent Minority	8
Mid-size Central City	Medium Percent Minority	8
Mid-size Central City	High Percent Minority	9
Urban Fringe of Mid-size Central City	Low Percent Minority	5
Urban Fringe of Mid-size Central City	Medium Percent Minority	4
Urban Fringe of Mid-size Central City	High Percent Minority	5
Large/Small Town	Low Percent Minority	11
Large/Small Town	Medium Percent Minority	12
Large/Small Town	High Percent Minority	10
Rural	Low Percent Minority	11
Rural	Medium Percent Minority	10
Rural	High Percent Minority	10
		103
NORTH DAKOTA (Small School Cluster Type 3 - Stratified)		
Large Schools		
Mid-size Central City and Urban Fringe	None	10
Large/Small Town	None	13
Rural	None	31
Small Schools		
None	None	19
		73
OHIO (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	9
Large/Mid-size Central City	Medium Percent Minority	9
Large/Mid-size Central City	High Percent Minority	9
Urban Fringe of Large/Mid-size Central City	None	34
Large/Small Town	None	23
Rural	None	21
		105

Table 3-4 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 8

Urbanization	Minority	Schools in Strata
OKLAHOMA (Small School Cluster Type 3 - Stratified)		
Large Schools		
Large/Mid-size Central City	Low Percent Minority	10
Large/Mid-size Central City	Medium Percent Minority	8
Large/Mid-size Central City	High Percent Minority	7
Urban Fringe of Large/Mid-size Central City	None	12
Large/Small Town	None	35
Rural	None	25
Small Schools		
None	None	10
		107
PENNSYLVANIA (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	6
Large/Mid-size Central City	Medium Percent Minority	6
Large/Mid-size Central City	High Percent Minority	6
Urban Fringe of Large/Mid-size Central City	None	33
Large/Small Town	None	35
Rural	None	16
		102
RHODE ISLAND (Small School Cluster Type 1 - None)		
Large Central City	Low Percent Minority	2
Large Central City	Medium Percent Minority	3
Large Central City	High Percent Minority	1
Mid-size Central City	None	4
Urban Fringe of Large/Mid-size Central City	None	27
Large/Small Town and Rural	None	12
		49
SOUTH CAROLINA (Small School Cluster Type 2 - Geographic)		
Mid-size Central City	Low Percent Minority	7
Mid-size Central City	Medium Percent Minority	5
Mid-size Central City	High Percent Minority	5
Urban Fringe of Mid-size Central City	Low Percent Minority	9
Urban Fringe of Mid-size Central City	Medium Percent Minority	10
Urban Fringe of Mid-size Central City	High Percent Minority	10
Small Town	Low Percent Minority	13
Small Town	Medium Percent Minority	12
Small Town	High Percent Minority	14
Rural	Low Percent Minority	6
Rural	Medium Percent Minority	7
Rural	High Percent Minority	7
		105

Table 3-4 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 8

Urbanization	Minority	Schools in Strata
TENNESSEE (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	10
Large/Mid-size Central City	Medium Percent Minority	10
Large/Mid-size Central City	High Percent Minority	11
Urban Fringe of Large/Mid-size Central City	None	18
Large/Small Town	None	31
Rural	None	24
		104
TEXAS (Small School Cluster Type 2 - Geographic)		
Large Central City	Low Hispanic/Low Black	6
Large Central City	Low Hispanic/High Black	5
Large Central City	High Hispanic/Low Black	6
Large Central City	High Hispanic/High Black	6
Mid-size Central City	Low Percent Minority	8
Mid-size Central City	Medium Percent Minority	9
Mid-size Central City	High Percent Minority	8
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	6
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	7
Urban Fringe of Large/Mid-size Central City	High Percent Minority	6
Large/Small Town and Rural	Low Percent Minority	12
Large/Small Town and Rural	Medium Percent Minority	13
Large/Small Town and Rural	High Percent Minority	12
		104
UTAH (Small School Cluster Type 2 - Geographic)		
Mid-size Central City	None	21
Urban Fringe of Mid-size Central City	None	37
Rural	None	12
Large/Small Town	None	15
		85
VIRGINIA (Small School Cluster Type 2 - Geographic)		
Mid-size Central City	Low Percent Minority	11
Mid-size Central City	Medium Percent Minority	10
Mid-size Central City	High Percent Minority	8
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	11
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	12
Urban Fringe of Large/Mid-size Central City	High Percent Minority	12
Small Town	Low Percent Minority	5
Small Town	Medium Percent Minority	6
Small Town	High Percent Minority	6
Rural	Low Percent Minority	8
Rural	Medium Percent Minority	9
Rural	High Percent Minority	7
		105

Table 3-4 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 8

Urbanization	Minority	Schools in Strata
VIRGIN ISLANDS (Small School Cluster Type 1 - None)		
	Low Percent Minority	3
	Medium Percent Minority	2
	High Percent Minority	1
		6
WEST VIRGINIA (Small School Cluster Type 2 - Geographic)		
Mid-size Central City	None	14
Urban Fringe of Mid-size Central City	None	12
Rural	None	44
Large/Small Town	None	33
		103
WISCONSIN (Small School Cluster Type 2 - Geographic)		
Mid-size Central City	Low Percent Minority	15
Large/Mid-size Central City	Medium Percent Minority	12
Urban Fringe of Large/Mid-size Central City	None	17
Large/Small Town	None	33
Rural	None	28
		105
WYOMING (Small School Cluster Type 3 - Stratified)		
Large Schools		
Mid-size Central City	None	4
Urban Fringe of Mid-size Central City	Low Percent Minority	1
Urban Fringe of Mid-size Central City	Medium Percent Minority	1
Urban Fringe of Mid-size Central City	High Percent Minority	1
Small Town	None	26
Rural	None	16
Small Schools		
None	None	6
		55

Case 2: Urbanization strata with more than 10 percent Black students or 7 percent Hispanic students, but not more than 20 percent of both, were stratified by ordering percent minority enrollment within the urbanization classes and dividing the schools into three groups with about equal numbers of students per minority group.

Case 3: In urbanization strata with more than 20 percent of both Black and Hispanic students, minority strata were formed as follows. The minority group with the higher percentage gave the primary stratification variable; the remaining group gave the secondary stratification variable. Within urbanization class, the schools were sorted based on the primary stratification variable and divided into two groups of schools containing approximately equal numbers of students. Within each of these two groups, the schools were sorted by the secondary stratification variable and subdivided into two subgroups of schools containing approximately equal numbers of students. As a result, within urbanization strata there were four minority groups, low Black/low Hispanic, low Black/high Hispanic, high Black/low Hispanic, and high Black/high Hispanic.

Tables 3-3 and 3-4 provide information on minority stratification for the participating states, respectively for fourth and eighth grade.

3.4.4 Median Household Income

The median household income variable was not used as a prime stratification variable because the available income data were not up to date (i.e., they were based on the 1980 census. Instead, median household income was used as a sorting variable at the final state of stratification. Prior to the selection of the school samples, the schools were sorted by urbanization, then by minority classes within urbanization in a serpentine order, in which the sort alternated between descending and ascending order within each group. This meant that adjacent schools on the list were generally similar with regard to either urbanization or minority enrollment, and often to both. Within minority class, the schools were sorted, in serpentine order, by the median household income. This final stage of sorting resulted in implicit stratification of median income. The data on median household income, which were obtained from Donnelly Marketing Information Services, were related to the ZIP code area in which the school was located. The data are derived from the 1980 census, but expressed in 1985 dollars.

3.4.5 Schools With Fewer Than 20 Students

Schools with fewer than 20 students were collapsed with other schools to form a sampling unit of at least 20 students. The two methods used to collapse small schools are referred to as geographic and stratified grouping. This collapsing was done separately for fourth and eighth grade.

Geographic Grouping. If the number of small schools in the state was less than 20 percent, and the number of students in these small schools accounted for less than 1 percent of the total state grade enrollment, then each school was combined with a school close by geographically until the cluster contained at least 20 students.

Cluster level values for enrollment, urbanization, minority, income variables, and selection probabilities was equal to the corresponding values of the school in the cluster with largest enrollment.

Stratified Grouping. In states with a larger number of small schools (Cluster Type 3 states), schools were stratified into two groups. One group contained schools with fewer than 20 students, the other group contained schools with 20 or more students. The schools in the first group were clustered in the following manner. The schools were ordered from smallest to largest, then the largest school was matched with the smallest school. If this cluster contained 20 or more students, it was complete. If the total cluster enrollment was 19 or smaller, the next smallest school was added. This continued until the sum of the enrollment was at least 20. We proceeded to form the next cluster with the next largest and smallest school in the same manner. If, after forming all the clusters, there remained a cluster with fewer than 20 students, it was combined with the previous cluster.

The enrollment value assigned to a cluster was equal to the sum of enrollments of the schools in that cluster. The minority value assigned to the cluster was equal to the weighted average of the proportion of minority for schools in the cluster where the weight was the fourth/eighth-grade enrollment. The cluster level income value was the median income of the school with the largest enrollment. No urbanization value was desired for clusters of schools. Also, no selection probability was derived for clusters of schools since they were selected with equal probability.

Tables 3-3 and 3-4 show the type of stratification used for small schools within the participating states, for fourth- and eighth-grade samples.

3.5 SCHOOL SAMPLE SELECTION FOR THE 1992 TRIAL STATE ASSESSMENT

3.5.1 Control of Overlap of School Samples for National Educational Studies

The issue of school sample overlap has been relevant in all rounds of NAEP in recent years, but no more so than in 1992. NAEP collected data nationally from a number of distinct samples at all three age classes, while state assessments were conducted at grades 4 and 8. At the same time, the U.S. Department of Education conducted the first phase followup to *Prospects: The National Longitudinal Study of Chapter I Children* (Abt Associates, 1991), for which a sample of districts was selected prior to the 1992 Trial State Assessment sample selection. This study involved substantial student assessment at grades 4 and 8.

To avoid undue burden on individual schools, NAEP developed a policy for 1992 of avoiding overlap of school samples from different studies for the same age class. This was to be achieved without unduly distorting the resulting samples by introducing bias or substantial variance. Thus, at grade 8 for example, the school samples for the national samples, the state samples, and the Prospects samples were selected to contain different schools, to the extent feasible. Besides generally controlling overlap within grade, distinct schools were selected for the fourth- and eighth-grade state assessment samples within a state to the extent feasible. The procedure used was an extension of the method proposed by Keyfitz (1951). The general approach is as follows.

Consider as an example the selection of samples for the Trial State Assessment eighth-grade sample. At the time of drawing the NAEP samples, the identities of the Prospects sample schools were not known. Since the selected districts, and district selection probabilities for all districts, were known, this information was used to control sample overlap. For each school in the frame for the national and state NAEP samples, there was a flag, C, indicating whether (C=1) or not (C=0) the district containing the school was included in the Prospects sample, and a Prospects district selection probability, $P_c = P(C=1)$.

In controlling overlap between NAEP state and national sample school selections, we used national school selection probabilities that were conditional on the selection of national sample PSUs (i.e., the school-within-PSU selection probabilities). This meant that in selecting the state samples, in those states where there was no PSU selection for the national samples no adjustments were needed to account for the selection of national NAEP samples (which might have selected schools within that state but, in fact, did not). This procedure of conditioning on the selection of PSUs also recognizes the impact of the heavy within-PSU sampling in noncertainty PSUs in some states, even though the unconditional probabilities of selection for such schools in the national samples were quite low. In other words, conditioning on the national PSU sample reduces the variance of the state samples, although it leads to a greater degree of sample overlap than if unconditional national selection probabilities had been used in the procedure for controlling overlap between state and national samples.

Let N = 1 if the school is selected in the national sample; let N = 0 otherwise. Let $P_N = P(N = 1)$. Thus, $P_N = 0$ for schools not located within a selected national sample PSU. Let π_s denote the expected number of times a school is to be selected for the state eighth-grade sample. The actual number of times that a school will be selected for the sample with the systematic sampling procedure used is equal to π_s if π_s is an integer, or to one of the two integers closest to π_s if π_s is not an integer. Large schools within a state may be selected up to three times; that is, π_s can be as great as 3 for some schools. The sample size of students to be drawn within the school is proportional to the number of times the school is selected. Schools to be included with certainty in the state sample ($\pi_s \geq 1$) are not subject to overlap control, as such schools are self-representing in the state sample. Excluding such schools on a random basis would add undue variance to the state estimates.

For $\pi_s < 1$, π_s denotes a true probability of selection for the school. Where possible, schools in districts selected for the Prospects study were excluded provided that the Prospects' district selection probability, P_c, fell below a constant, k_s, that varied from state to state. In small states, where it is important to include all eligible schools in the state sample, k_s was set to zero. The variable C indicates whether (C=1) or not (C=0) the district was included in the Prospects sample. For actually drawing the state samples, a conditional expected number of selections, π_s^*, was derived for each school in the frame as follows:

$$\pi_s^* = \pi_s \qquad \text{if } \pi_s \geq 1$$

$$\pi_s/(1 - \phi_N) \qquad \text{if } \pi_s < 1,\ P_c > k_s, \text{ and } N = 0$$

$$\frac{\pi_s(P_N - \phi_N)}{P_N(1 - \phi_N)} \qquad \text{if } \pi_s < 1,\ P_C > k_p, \text{ and } N = 1$$

$$\pi_s/(1 - \nu_{NC}) \qquad \text{if } \pi_s < 1,\ P_C \le k_p, \text{ and } (C = 0 \text{ and } N = 0)$$

$$\frac{\pi_s(P_N + P_C - \nu_{NC})}{(P_N + P_C)(1 - \nu_{NC})} \qquad \text{if } \pi_s < 1,\ P_C \le k_p, \text{ and } (C = 1 \text{ or } N = 1)$$

where $\phi_N = \min(P_N, 1 - \pi_s)$ and $\nu_{NC} = \min(P_N + P_C, 1 - \pi_s)$. The values of π_s^* are conditional on the selection of districts for the Prospects sample and PSUs for the national NAEP samples.

This procedure in general gave state NAEP conditional selection probabilities that are smaller than the unconditional selection probabilities for schools located in Prospects selected districts, and for schools selected for the national sample. The relative chance of selection in the state sample for a school selected in either of these other two samples, compared to its chance of selection in the state sample if not selected for either of the other samples, is $(P_N - \phi_N)/P_N$ if $P_C > k_p$ and $(P_N + P_C - \nu_{NC})/(P_N + P_C)$ if $P_C \le k_p$. If P_N, P_C, and π_s are all relatively small, then $P_N + P_C - \nu_{NC} = 0$, so that there was no chance of selecting the school for the state sample if it is in the national sample or in a Prospects district selection. The expected number of times a school was selected in the state sample, conditional on the national PSU sample but unconditional on the national school sample selection within PSUs and the selection of districts for the Prospects sample, is given by π_s, as desired. This follows from the above formulation of π_s^* and the fact that $P(C=1 \text{ or } N=1)$ equals $P_N + P_C$ when $P_C \le k_p$ since there is no overlap of NAEP national sample selected schools and Prospects selected districts in this case. The quantity π_s is used as the basis for weighting the schools, and hence students, in the state samples.

To illustrate the use of these expressions in drawing the state sample, suppose that $P_C > k_p$ (or $P_C=0$) so that we are concerned only with controlling overlap with the national sample. Suppose that $\pi_s = 0.3$ and $P_N = 0.25$. Then $\phi_N = P_N = 0.25$, and $\pi_s^* = 0.4$ if the school is not selected for the national sample. Thus in this case the school is selected to conduct a single assessment session of about 30 students with probability 0.4. Since $\phi_N = P_N$, $\pi_s^* = 0$ if the school is selected for the national sample. Thus there is no chance that this school will be selected for both national and state samples. Integrating over the national sampling process gives the required unconditional state selection probability of 0.3 ($= 0.4 * (1 - 0.25) + 0 * 0.25$).

3.5.2 Selection of Schools in Small States (Cluster Type 1 States)

For states with small numbers of schools, and no or very few small schools, all schools were included in the sample with certainty. In the fourth grade, all the eligible fourth-grade schools in the District of Columbia, Delaware, Guam, and the Virgin Islands were taken into the sample with certainty. In the eighth grade, all the eligible schools were taken from the District of Columbia, Delaware, Hawaii, Rhode Island, Guam, and the Virgin Islands.

3.5.3 States with Geographic Clustering of Small Schools (Cluster Type 2 States)

Clusters were sorted by urbanization, minority strata (which varied by state and urbanization level), and median income. A systematic sample of clusters was then selected for each state with probability proportionate to size, where size was equal to the estimated grade enrollment within the school, so as to achieve the desired student sample size of 3,150 for the eighth grade and 6,300 for the fourth grade.

Up to three sessions (90 students) were selected from each school to more efficiently represent the large schools in the eighth-grade sample. The fourth-grade sample selected two sessions from larger schools (those with more than 20 students), one for reading and one for mathematics assessments.

Following the selection of clusters, there was some thinning of small schools. The purpose of thinning was to give students in small schools (enrollment less than 20) approximately the same chance of selection as those from larger schools. In addition, thinning of small schools controlled the number of schools in the sample to be close to the desired number, and thereby controlling the cost of data collection. All small schools in a cluster were retained in the sample with probability P_s/P_c, where P_s was the probability of selection of the small school and P_c was the probability of selection of the cluster.

Table 3-5 shows the distribution of selected schools in the participating states.

3.5.4 States with Stratification of Small Schools (Cluster Type 3 States)

As described above, clusters were sorted by urbanization, minority strata (which varied by state and urbanization level), and median income within the two size clusters. Small school clusters were selected systematically with equal probability, and large schools were sampled systematically with probability proportionate to size, so as to achieve the desired student sample size of 3,150 for the eighth grade and 6,300 for the fourth grade.

Similar to Cluster Type 2 states, up to three eighth-grade sessions were selected within each school to more efficiently represent the larger schools in the sample. For the fourth-grade samples, each selected school was chosen for one reading and one mathematics session except for schools with fourth-grade enrollment of fewer than 20, which were assigned only a single session.

Table 3-5 shows the distribution of selected schools in the participating states.

3.5.5 Overlap of School Samples

As stated in section 3.5.1, the sample design for eighth-grade schools minimized, to the extent feasible, the chances of selecting eighth-grade schools in the 1992 national NAEP and the Prospects survey. Furthermore, the fourth-grade state samples were selected such that the number of schools in each state selected for both fourth- and eighth-grade samples were minimized to the extent feasible.

Table 3-5
Distribution of Sample Sizes by School Size, with Corresponding Overlap Between Grades

State	Number of Small* Schools Sampled for...			Number of Other Schools Sampled for...		
	4th Only	8th Only	4th & 8th	4th Only	8th Only	4th & 8th
Alabama	2	2	0	113	105	0
Arizona	6	6	0	106	102	0
Arkansas	7	3	0	119	98	0
California	8	0	0	108	105	0
Colorado	15	8	0	114	105	0
Connecticut	1	0	0	114	101	0
Delaware	0	0	2	50	24	2
District of Columbia	3	0	0	108	28	7
Florida	1	0	0	106	105	0
Georgia	0	0	0	109	104	0
Guam	0	0	0	21	6	0
Hawaii	1	2	0	94	43	12
Idaho	15	8	0	115	74	0
Indiana	2	0	0	116	105	0
Iowa	14	2	0	129	106	0
Kentucky	4	5	0	122	105	0
Louisiana	7	0	0	111	108	0
Maine	40	10	0	122	87	3
Maryland	2	0	0	109	104	0
Massachusetts	1	0	0	122	105	0
Michigan	2	0	0	114	105	0
Minnesota	4	0	0	116	104	0
Mississippi	1	0	0	110	102	0
Missouri	14	6	0	116	101	0
Nebraska	79	42	0	121	79	0
New Hampshire	25	3	0	115	75	0
New Jersey	4	1	0	119	106	0
New Mexico	12	7	0	110	86	0
New York	1	2	0	109	105	0
North Carolina	4	1	0	113	105	0
North Dakota	46	23	0	118	55	1
Ohio	2	0	0	116	105	0
Oklahoma	26	12	0	118	98	0
Pennsylvania	0	0	0	118	104	0
Rhode Island	0	0	1	108	47	6
South Carolina	2	0	0	111	105	0
Tennessee	7	1	0	115	105	0
Texas	5	2	0	109	105	0
Utah	6	1	0	106	87	0
Virginia	3	0	0	111	106	0
Virgin Islands	1	0	0	22	5	1
West Virginia	26	1	0	140	106	0
Wisconsin	14	2	0	120	105	0
Wyoming	61	3	12	120	47	2

---

*Small school denotes a school with fewer than 20 students.

BEST COPY AVAILABLE

Table 3-5 shows the overlap of fourth- and eighth-grade schools in participating states. Table 3-6 shows the number of schools selected in both the national and state assessments.

3.5.6 New School Selection

A district-level file was constructed from the aggregate of the fourth- and eighth-grade school frame. The file was divided into a small districts file, consisting of those districts in which there were at most two schools on the aggregate frame but no more than one fourth- and one eighth-grade school. The remainder of districts were denoted as "large" districts.

All new eligible schools coming from "small" districts (those with at most one grade 4 and one grade 8 school) that had a school selected in the regular sample for that grade were included in the sample and treated as belonging to the same cluster as the original selection from that district.

A sample of "large" districts was drawn in each state. All districts were selected in Delaware, the District of Columbia, Guam, Hawaii, and the Virgin Islands. The remainder of the states in the file of "large" districts (eligible for sampling) was divided in two files within each state; two districts were selected with equal probability among the districts with combined enrollment of about 20 percent of the state enrollment.

From the rest of the file, eight districts per state were selected with probability proportional to enrollment. The selected districts were then sent a listing of all their schools that appeared on the QED sampling frame, and were asked to provide information about the new schools not included in the QED frame. These listings, provided by selected districts, were used as sampling frames for selection of new schools.

The eligibility of a school was determined based on the grade span. A school was classified as "new" if the school was eligible for sampling based on its grade span but not included in the QED frame, or if the changes of grade span were such that the school status changed from ineligible to eligible. The average grade enrollment for these schools was set to the average grade enrollment before the grade span change. The schools found eligible for sampling due to the grade span change were added to the corresponding grade frame.

Similar to the main sample, we assigned the following measure of size to eighth-grade schools to produce self-weighting samples of students:

$$\begin{cases} 30 & \text{if eighth-grade enrollment} < 35 \\ \textit{eighth-grade enrollment} & \text{otherwise} \end{cases}$$

The probability of selecting a school was $\min\left[\dfrac{\text{sampling rate} \bullet \text{measure of size}}{\text{P(district)}}, 1\right]$,

where P(district) was the probability of selection of a district and the sampling rate was the rate used for the particular state in the selection of the original sample of schools.

Table 3-6
Number of Schools Selected for Both National and State Samples, by State

State	Both Grade 4	Both Grade 8	Different Grades	Total
Alabama	0	1	2	3
Arizona	0	1	3	4
Arkansas*	1	2	5	7
California	0	0	0	0
Colorado	0	0	0	0
Connecticut	0	4	3	7
Delaware	0	0	0	0
District of Columbia	0	1	0	1
Florida	0	2	1	3
Georgia	0	1	0	1
Hawaii	0	0	0	0
Idaho	0	0	0	0
Indiana	0	2	1	3
Iowa	0	1	0	1
Kentucky	0	3	1	4
Louisiana	0	2	7	9
Maine	0	0	0	0
Maryland	0	2	0	2
Massachusetts	0	0	0	0
Michigan	0	0	0	0
Minnesota	0	1	1	2
Mississippi	0	0	0	0
Missouri	0	1	0	1
Nebraska*	0	8	5	12
New Hampshire	0	0	0	0
New Jersey	0	0	1	1
New Mexico	0	4	4	9
New York	1	1	0	1
North Carolina	0	0	0	0
North Dakota	0	0	0	0
Ohio	0	0	1	1
Oklahoma	0	1	1	2
Pennsylvania	0	1	2	3
Rhode Island	0	3	0	3
South Carolina	0	1	1	2
Tennessee	0	0	2	2
Texas	0	0	0	0
Utah	0	0	0	0
Virginia	0	0	0	0
West Virginia	0	2	2	4
Wisconsin	0	2	2	4
Wyoming	1	3	3	7
TOTAL	3	50	48	101

---

*One school was selected for State Grade 8 and National Grades 8 and 12.

The probability of selection of a school for the fourth-grade sample was the same as the eighth grade, with the measure of size for the fourth-grade schools being

$$\begin{cases} 60 & \text{if enrollment} \leq 70 \\ \textit{enrollment} & \text{if enrollment} > 70 \end{cases}$$

The selection of the fourth- and eighth-grade sample was independent; thus, no selection probability adjustments was needed from one grade selection to the other. In each state, the sampling rates used for the main sample of fourth- and eighth-grade schools were used to select the new schools for the fourth- and eighth-grade samples, respectively.

Tables 3-7 and 3-8 show the number of new schools coming from the "large" and "small" districts for the fourth- and eighth-grade samples.

3.5.7 Assigning Subject Session Types at Grade 4

In the interest of sampling efficiency it was desirable that each of the two subjects assessed at grade 4, reading and mathematics, be administered in as large a subset of the sampled schools as possible. On the other hand it was unreasonable to expect very small schools to conduct two different sessions with half of the eligible students in each. To satisfy these two requirements the following procedure was used.

If, according to the information on the frame, the school had an enrollment of 21 or more grade 4 students, the school was assigned initially to conduct both mathematics and reading sessions, with half of the selected students being assigned to a mathematics session, and half to a reading session (see section 3.6 for a description of the student sampling process). This varied only in Guam, where all students took both assessment types.

If, according to the frame data, the school enrollment was 20 or fewer, the school was randomly assigned to conduct either a mathematics or a reading session. The assignment was systematic, based on the ordering of the clusters for sample selection, with random ordering of selected schools within clusters.

If a school had two session types assigned initially, but was found at the time of drawing the student samples to have fewer than 21 eligible students, the school was randomly assigned to conduct only one of the two session types, with each type being chosen with probability 0.5. This assignment was independent from school to school. Thus a school was to conduct a single session type if either its frame or its actual enrollment for grade 4 was 20 or fewer; a school was to conduct both session types if both its frame and actual enrollments exceeded 20.

3.5.8 Designating Monitor Status

Within each state, random equivalent half samples of schools were assigned to be monitored or unmonitored. The details of the implementation of the monitoring process in the

Table 3-7
Distribution of New Schools Coming from "Large" and "Small" Districts in the Fourth-grade Sample

State	Number of New Schools	
	"Large" Districts	"Small" Districts
Alabama	-	-
Arizona	-	-
Arkansas	1	-
California	2	1
Colorado	2	-
Connecticut	-	-
Delaware	3	-
District of Columbia	1	-
Florida	5	-
Georgia	1	-
Guam	-	-
Hawaii	1	-
Idaho	-	-
Indiana	1	-
Iowa	-	-
Kentucky	2	-
Louisiana	-	-
Maine	-	-
Maryland	2	-
Massachusetts	-	1
Michigan	1	-
Minnesota	1	-
Mississippi	1	-
Missouri	1	-
Nebraska	-	-
New Hampshire	-	-
New Jersey	1	-
New Mexico	-	-
New York	-	-
North Carolina	4	-
North Dakota	1	-
Ohio	5	-
Oklahoma	-	-
Pennsylvania	1	-
Rhode Island	-	-
South Carolina	1	-
Tennessee	2	-
Texas	-	-
Utah	-	1
Virginia	5	-
Virgin Islands	-	-
West Virginia	1	-
Wisconsin	1	-
Wyoming	-	-

BEST COPY AVAILABLE

Table 3-8
Distribution of New Schools Coming from "Large" and "Small" Districts in the Eighth-grade Sample

State	Number of New Schools	
	"Large" Districts	"Small" Districts
Alabama	-	-
Arizona	1	-
Arkansas	-	-
California	1	1
Colorado	-	-
Connecticut	-	-
Delaware	2	-
District of Columbia	2	-
Florida	2	-
Georgia	2	-
Guam	-	-
Hawaii	-	-
Idaho	-	-
Indiana	2	-
Iowa	-	1
Kentucky	2	-
Louisiana	1	-
Maine	-	-
Maryland	-	-
Massachusetts	-	4
Michigan	3	-
Minnesota	-	-
Mississippi	-	-
Missouri	-	-
Nebraska	1	-
New Hampshire	-	-
New Jersey	1	-
New Mexico	-	-
New York	1	-
North Carolina	2	-
North Dakota	1	-
Ohio	5	-
Oklahoma	-	-
Pennsylvania	2	1
Rhode Island	1	2
South Carolina	-	-
Tennessee	-	-
Texas	-	-
Utah	-	-
Virginia	2	-
Virgin Islands	-	-
West Virginia	1	-
Wisconsin	2	-
Wyoming	2	-

field are given in Chapter 4. The purpose of monitoring a random half of the schools was to ensure that the procedures were being followed throughout each state by the school and district personnel administering the assessments, and to provide data adequate for assessing whether there was a significant difference in assessment results between monitored and unmonitored schools within each state.

The following procedure was used to determine the sample of schools to be monitored. The initially selected clusters were sorted in the order in which they were systematically selected (see sections 3.5.2 to 3.5.4). New schools from "large" districts added to the sample (see section 3.5.6) were treated as single school clusters, and were added to the end of the list in random order. The sorted clusters were then paired, and one member of each pair was assigned at random, with probability 0.5, to be monitored. The assignment was independent across pairs. If there was an odd number of clusters, the last cluster was assigned, with 0.5 probability, to be monitored.

If a cluster was designated to be monitored, all selected schools within the cluster (after thinning of small schools from multiple school clusters in Cluster Type 2 states; see section 3.5.3) were assigned to be monitored. For the grade 4 samples, this procedure, in combination with the procedure for assigning schools to subjects (see section 3.5.7), ensured that for every pair of clusters for each subject at least one school would be monitored and at least one would not.

In the territories of Guam and the Virgin Islands, there were few schools in each sample, and large samples of students (that is, all of the students enrolled) were drawn from each school. In these jurisdictions the monitoring assignment was done at the level of the physical assessment session, rather than at the cluster level. After establishing in each school the number of sessions to be conducted, alternate sessions were designated to be monitored, with the first session assigned at random. Thus all schools contained some monitored and some unmonitored sessions.

3.5.9 Substitutes

A substitute school was selected for each selected school containing eligible students, for which school nonparticipation was established by the state coordinator as of November 1, 1991. The process of selecting a substitute for a school involved identifying the most similar school in terms of the following characteristics: urbanization, percent Black enrollment, percent Hispanic enrollment, fourth-grade (or eighth-grade, as applicable) enrollment, and median income. To identify candidates for substitution, a set of schools was found that provided reasonable matches with regard to fourth/eighth-grade enrollment, and percent Black and Hispanic enrollment. From among this set a match was selected, considering all five characteristics. Schools in the National Assessment sample and those in the Prospects study were avoided in the selection of substitutes, where possible. Furthermore, the substitute was selected from the same district, wherever possible, to avoid placing the burden of replacing a refusing school from one district on another district. This was often not possible, however, as in the majority of cases the decision not to participate was made at the district level.

In the cases where no suitable substitute could be found among those schools not sampled (most often because all or most schools were included in the original sample), a school already in the sample conducted a double session, of which one session served as a substitute for students in the refusing school. The same criteria were applied in selecting the schools that conducted double sessions; that is, a reasonable match was found based on grade enrollment, percent of Black and Hispanic enrollment, median income, and urbanization.

Tables 3-9 and 3-10 include information about the number of substitutes provided for each grade and in each state. Of the 44 states participating, 27 were provided with at least one substitute. Among states receiving no substitutes, the majority had 100 percent participation from the original sample. In a few cases, however, refusals did occur after the November 1 deadline. The number of substitutes provided to a state ranged from 0 to 59 in the fourth grade and 0 to 43 in the eighth grade. A total of 591 substitutes were selected for the fourth-grade sample, 23 of which were double session substitutes. A total of 460 substitutes were selected for the eighth-grade sample, 75 of which were double session substitutes. Some states did not attempt to solicit participation from the substitute schools provided, as they considered the timing too late to seek cooperation from schools not previously notified about the assessment. In quite a few cases the originally selected school agreed to cooperate after a substitute was selected and had agreed to participate (in which case the substitute school data were discarded).

Tables 3-11 and 3-12 show the number of schools in the fourth- and eighth-grade samples for the mathematics assessment, together with school response rates observed within participating states. Refer to the Trial State Assessment report entitled *School and Student Participation Rates for the Mathematics Assessment and Guidelines for Sample Participation*, September 1992, for an analysis of participation rates. The tables also show the number of substitutes in each state that were associated with a nonparticipating original school selection, and the number of those that participated.

3.6 STUDENT SAMPLE SELECTION

Schools initially sent a complete list of students to a central location in November 1991. Schools were not asked to list students in any particular order, but were asked to implement checks to ensure that all fourth/eighth-grade students were listed. Based on the total number of students on this list, called the Student Listing Form, sample line numbers were generated for student sample selection. To generate these line numbers, the sampler entered the number of students on the form and the number of mathematics and reading sessions into a calculator that had been programmed with the sampling algorithm. The calculator generated a random start that was used to systematically select the student line numbers (30 per session). To compensate for new enrollees not on the Student Listing Form, extra line numbers were generated for a supplemental sample of new students. All students were selected in those schools with grade enrollment size of up to 10 percent more than the required sample size of students. This sample design was intended to give each student within the state approximately the same chance of selection.

The states where all schools were selected with certainty (Cluster Type 1 states) were treated differently. For the fourth-grade sample in Delaware and the District of Columbia, 120 students were selected, where possible. If the enrollment was lower than 120, all of the students were taken. In the territories, all of the fourth-grade students were included in the sample. In the six states where all schools were selected at grade 8, up to 90 students were selected for the sample from each school, depending on the school size.

Table 3-9
Substitute School Counts for Grade 4

State	Double Session Substitutes	Regular Substitutes	Total
Alabama	2	27	29
Arkansas	0	13	13
California	0	16	16
Idaho	0	24	24
Indiana	0	28	28
Kentucky	0	3	3
Maine	3	53	56
Maryland	0	1	1
Massachusetts	0	15	15
Michigan	0	20	20
Minnesota	1	15	16
Mississippi	0	2	2
Missouri	0	9	9
Nebraska	0	59	59
New Hampshire	0	42	42
New Jersey	0	53	53
New Mexico	2	32	34
New York	0	28	28
North Carolina	0	5	5
North Dakota	1	46	47
Ohio	0	27	27
Oklahoma	0	15	15
Pennsylvania	0	17	17
Rhode Island	14	2	16
South Carolina	0	2	2
Tennessee	0	8	8
Texas	0	6	6
TOTAL	23	568	591

Table 3-10
Substitute School Counts for Grade 8

State	Double Session Substitutes	Regular Substitutes	Total
Alabama	4	36	40
Arkansas	0	10	10
California	1	13	14
Idaho	7	6	13
Indiana	2	19	21
Kentucky	0	3	3
Maine	16	19	35
Maryland	3	7	10
Massachusetts	0	12	12
Michigan	0	23	23
Minnesota	3	14	17
Mississippi	0	1	1
Missouri	0	8	8
Nebraska	4	30	34
New Hampshire	4	11	15
New Jersey	0	43	43
New Mexico	17	7	24
New York	1	23	24
North Carolina	0	4	4
North Dakota	1	16	17
Ohio	0	23	23
Oklahoma	2	15	17
Pennsylvania	1	24	25
Rhode Island	8	0	8
South Carolina	0	4	4
Tennessee	1	9	10
Texas	0	5	5
TOTAL	75	385	460

Table 3-11
Distribution of the Grade 4 Mathematics School Sample by State

State	Weighted Percent School Participation		Number of Schools in the Original Sample			Number of Substitute Schools for Nonparticipating Originals		Total Number of Schools That Participated
	Before Substitution	After Substitution	Total	Not Eligible	Participated	Provided	Participated	
Alabama	74.91	97.04	113	3	81	27	25	106
Arizona	100.00	100.00	110	2	108	0	0	108
Arkansas	89.65	99.13	123	2	109	11	11	120
California	91.30	96.98	115	3	101	7	7	108
Colorado	100.00	100.00	123	2	121	0	0	121
Connecticut	99.03	99.03	115	4	110	0	0	110
Delaware	92.15	92.15	56	6	44	0	0	44
Dist. of Columbia	98.82	98.82	114	5	107	0	0	107
Florida	100.00	100.00	111	1	110	0	0	110
Georgia	100.00	100.00	110	2	108	0	0	108
Guam	94.18	94.18	21	0	20	0	0	20
Hawaii	100.00	100.00	108	0	108	0	0	108
Idaho	83.57	96.62	120	0	98	21	17	115
Indiana	76.31	91.06	118	2	88	26	17	105
Iowa	100.00	100.00	132	4	128	0	0	128
Kentucky	92.85	95.65	124	3	115	3	3	118
Louisiana	100.00	100.00	113	4	109	0	0	109
Maine	56.51	71.23	142	2	75	43	22	97
Maryland	99.19	99.19	112	1	110	1	0	110
Massachusetts	86.51	96.64	123	4	103	12	11	114
Michigan	83.04	89.57	114	3	90	16	8	98
Minnesota	81.59	93.74	118	5	93	15	13	106
Mississippi	98.09	100.00	111	2	107	2	2	109
Missouri	89.25	97.06	120	7	101	9	9	110
Nebraska	79.88	87.25	157	6	109	36	11	120
New Hampshire	68.72	80.28	126	3	84	32	16	100
New Jersey	75.86	81.84	119	3	88	22	7	95
New Mexico	75.45	90.38	116	2	86	26	18	104
New York	77.74	83.35	107	0	83	21	7	90
North Carolina	95.15	99.09	118	2	111	5	5	116
North Dakota	73.06	89.84	133	3	97	30	19	116
Ohio	78.63	91.29	122	1	95	21	15	110
Oklahoma	86.14	98.06	129	3	111	14	13	124
Pennsylvania	84.27	95.40	116	0	99	17	12	111
Rhode Island	83.32	96.16	115	5	90	15	15	105
South Carolina	98.06	99.03	112	2	108	1	1	109
Tennessee	91.66	92.71	120	2	108	8	1	109
Texas	93.09	97.93	111	3	100	5	5	105
Utah	99.05	99.05	110	1	108	0	0	108
Virginia	98.99	98.99	116	4	111	0	0	111
Virgin Islands	100.00	100.00	24	0	24	0	0	24
West Virginia	100.00	100.00	147	6	141	0	0	141
Wisconsin	100.00	100.00	127	5	122	0	0	122
Wyoming	96.77	96.77	157	11	143	0	0	143

Table 3-12
Distribution of the Grade 8 Mathematics School Sample by State

State	Weighted Percent School Participation		Number of Schools in the Original Sample			Number of Substitute Schools for Nonparticipating Originals		Total Number of Schools That Participated
	Before Substitution	After Substitution	Total	Not Eligible	Participated	Provided	Participated	
Alabama	65.71	92.28	107	1	70	33	28	98
Arizona	98.73	98.73	109	5	103	0	0	103
Arkansas	89.44	97.22	101	1	89	10	8	97
California	93.34	98.10	107	2	98	6	5	103
Colorado	100.00	100.00	113	1	112	0	0	112
Connecticut	99.02	99.02	101	3	97	0	0	97
Delaware	100.00	100.00	30	2	28	0	0	28
Dist. of Columbia	100.00	100.00	37	2	35	0	0	35
Florida	100.00	100.00	107	4	103	0	0	103
Georgia	99.02	99.02	106	4	102	0	0	102
Guam	100.00	100.00	6	0	6	0	0	6
Hawaii	99.97	99.97	57	5	51	0	0	51
Idaho	84.78	91.06	82	1	67	11	6	73
Indiana	79.28	93.69	107	0	85	20	16	101
Iowa	99.06	99.06	109	3	105	0	0	105
Kentucky	96.25	98.13	112	6	102	3	2	104
Louisiana	100.00	100.00	109	8	101	0	0	101
Maine	62.15	84.20	100	0	60	31	20	80
Maryland	89.41	91.34	104	1	93	9	2	95
Massachusetts	83.32	95.14	109	7	85	12	12	97
Michigan	77.59	94.40	108	1	83	22	18	101
Minnesota	81.39	92.16	104	3	82	15	11	93
Mississippi	98.93	100.00	102	3	98	1	1	99
Missouri	92.16	99.02	107	1	98	7	7	105
Nebraska	75.19	85.23	122	10	73	34	12	85
New Hampshire	79.86	91.67	78	1	62	12	9	71
New Jersey	69.26	77.73	108	2	75	27	9	84
New Mexico	77.35	93.96	93	1	69	21	15	84
New York	80.57	83.48	108	4	84	19	3	87
North Carolina	94.30	98.10	108	3	99	4	4	103
North Dakota	78.37	96.78	80	6	55	16	15	70
Ohio	77.21	89.48	110	0	85	20	14	99
Oklahoma	81.77	98.29	110	3	88	17	17	105
Pennsylvania	80.78	94.23	107	2	84	20	14	98
Rhode Island	85.04	99.66	57	5	44	7	7	51
South Carolina	94.10	97.17	105	0	99	4	3	102
Tennessee	87.46	91.32	106	2	91	10	4	95
Texas	95.14	99.03	107	3	99	5	4	103
Utah	100.00	100.00	88	3	85	0	0	85
Virginia	97.17	97.17	108	2	103	0	0	103
Virgin Islands	100.00	100.00	6	0	6	0	0	6
West Virginia	100.00	100.00	108	4	104	0	0	104
Wisconsin	100.00	100.00	109	2	107	0	0	107
Wyoming	99.04	99.04	66	11	54	0	0	54

After the student sample was selected, the administrator at each school identified students who were incapable of taking the assessment because they were either disabled or unable to speak English. More details on the procedures for student exclusion are presented in the report on field procedures for the Trial State Assessment Program.

When the assessment was conducted in a given school, a count was made of the number of nonexcluded students who did not attend the session. If this number exceeded three students, the school was instructed to conduct a make-up session, to which were invited all students who were absent from the initial session.

Tables 3-13 and 3-14 provide the distribution of the fourth-grade and eighth-grade mathematics student samples and response rates by state.

Table 3-13
Distribution of the Grade 4 Mathematics Student Sample and Response Rates by State

State	Weighted Student Response Rate (Percent)	Number of Students			
		In Original Sample	Excluded from Sample	To Be Assessed	Actually Assessed
Alabama	95.39	2,903	127	2,729	2,605
Arizona	95.45	3,133	154	2,899	2,741
Arkansas	96.34	2,961	154	2,748	2,621
California	93.96	3,015	364	2,568	2,412
Colorado	95.33	3,244	166	3,050	2,906
Connecticut	95.88	2,959	196	2,713	2,600
Delaware	94.95	2,330	121	2,152	2,040
District of Columbia	93.15	2,914	255	2,577	2,399
Florida	94.94	3,267	273	2,982	2,828
Georgia	95.38	3,117	154	2,899	2,766
Guam	94.85	2,158	133	2,038	1,933
Hawaii	95.03	3,009	169	2,761	2,625
Idaho	96.89	2,983	102	2,871	2,784
Indiana	95.71	2,815	92	2,709	2,593
Iowa	95.97	3,001	98	2,883	2,770
Kentucky	95.75	2,970	99	2,824	2,703
Louisiana	95.03	3,113	122	2,938	2,792
Maine	94.95	2,161	124	2,022	1,893
Maryland	95.59	3,170	126	2,972	2,844
Massachusetts	95.23	2,942	219	2,678	2,549
Michigan	94.11	2,736	136	2,582	2,412
Minnesota	95.43	2,924	104	2,799	2,640
Mississippi	96.55	3,023	146	2,807	2,712
Missouri	95.82	2,778	117	2,621	2,509
Nebraska	95.59	2,602	122	2,444	2,327
New Hampshire	96.08	2,538	99	2,408	2,265
New Jersey	96.04	2,483	133	2,322	2,231
New Mexico	95.10	2,874	188	2,552	2,342
New York	95.81	2,545	127	2,387	2,284
North Carolina	95.35	3,144	122	3,022	2,884
North Dakota	96.49	2,312	45	2,269	2,193
Ohio	95.37	2,962	166	2,767	2,637
Oklahoma	84.45	2,936	215	2,682	2,254
Pennsylvania	95.62	3,015	112	2,868	2,740
Rhode Island	94.85	2,767	161	2,518	2,390
South Carolina	96.55	3,045	144	2,868	2,771
Tennessee	95.98	2,979	117	2,821	2,708
Texas	96.46	3,013	234	2,722	2,623
Utah	95.55	3,130	128	2,930	2,799
Virginia	95.26	3,105	163	2,926	2,786
Virgin Islands	97.11	952	24	932	905
West Virginia	95.60	3,068	134	2,914	2,786
Wisconsin	95.85	3,079	141	2,910	2,780
Wyoming	95.92	2,833	98	2,717	2,605

Table 3-14
Distribution of the Grade 8 Mathematics Student Sample and Response Rates by State

State	Weighted Student Response Rate (Percent)	Number of Students			
		In Original Sample	Excluded from Sample	To Be Assessed	Actually Assessed
Alabama	95.43	3,011	165	2,748	2,522
Arizona	92.71	3,089	178	2,812	2,617
Arkansas	94.01	2,978	176	2,717	2,556
California	91.85	3,101	246	2,763	2,516
Colorado	93.18	3,199	136	3,006	2,799
Connecticut	93.73	3,029	192	2,783	2,613
Delaware	91.83	2,220	97	2,098	1,934
District of Columbia	84.95	2,517	225	2,137	1,816
Florida	90.67	3,073	199	2,812	2,549
Georgia	92.96	3,011	137	2,787	2,589
Guam	89.74	1,734	72	1,667	1,496
Hawaii	89.69	2,904	142	2,724	2,454
Idaho	94.59	2,936	91	2,799	2,615
Indiana	94.29	3,000	140	2,820	2,659
Iowa	95.29	3,133	129	2,959	2,816
Kentucky	95.52	3,087	135	2,883	2,756
Louisiana	92.34	3,028	120	2,794	2,582
Maine	93.22	2,838	124	2,698	2,520
Maryland	92.00	2,803	128	2,605	2,399
Massachusetts	93.55	2,909	217	2,623	2,456
Michigan	93.65	3,020	184	2,793	2,616
Minnesota	94.22	2,758	92	2,619	2,471
Mississippi	94.75	2,958	207	2,636	2,498
Missouri	94.97	2,984	128	2,815	2,666
Nebraska	95.51	2,543	108	2,392	2,285
New Hampshire	93.56	2,958	156	2,755	2,582
New Jersey	94.06	2,506	169	2,307	2,174
New Mexico	92.98	3,041	163	2,780	2,561
New York	91.90	2,581	193	2,347	2,158
North Carolina	94.25	3,071	102	2,936	2,769
North Dakota	95.89	2,513	63	2,418	2,314
Ohio	92.71	2,942	177	2,732	2,535
Oklahoma	79.99	2,934	184	2,710	2,141
Pennsylvania	94.16	2,964	127	2,806	2,640
Rhode Island	92.52	2,481	119	2,289	2,120
South Carolina	93.51	3,057	174	2,808	2,625
Tennessee	94.10	2,838	137	2,644	2,485
Texas	93.56	3,048	205	2,794	2,614
Utah	93.74	3,124	141	2,910	2,726
Virginia	94.28	3,091	153	2,872	2,710
Virgin Islands	92.38	1,708	86	1,601	1,479
West Virginia	94.48	3,097	178	2,843	2,690
Wisconsin	93.68	3,165	130	3,002	2,814
Wyoming	94.78	2,743	107	2,576	2,444

BEST COPY AVAILABLE

Chapter 4

STATE AND SCHOOL COOPERATION AND FIELD ADMINISTRATION

Nancy Caldwell

Westat, Inc.

4.1 OVERVIEW

By volunteering to participate in the Trial State Assessment and in the field test that preceded it, each state assumed responsibility for securing the cooperation of the schools sampled by NAEP. The participating states were responsible for the actual administration of the 1992 Trial State Assessment at the school level. For the field test in 1991, however, individual states could choose to have NAEP administer the entire program. This chapter describes state and school cooperation and field administration procedures for both the field test and the 1992 program. Section 4.2 presents information on the field test in 1991, while section 4.3 focuses on the 1992 Trial State Assessment.

4.2 THE FIELD TEST

4.2.1 Conduct of the Field Test

In preparation for the 1992 state and national assessment programs, a field test of the forms, procedures, and booklet items was held in early 1991. The field test also gave states an opportunity to learn about their responsibilities for the new aspects of the Trial State Assessment.

In June 1990, letters were sent from the U.S. Department of Education to all Chief State School Officers inviting them to participate in the field test of materials and procedures for 1992. Since the fourth grade had not been assessed as part of the Trial State Assessment before, states were given the option of conducting the field test themselves this grade. Only states that had not participated in the 1990 assessment at the eighth grade were given the option of conducting the field test themselves. Otherwise, NAEP staff were to conduct the field test.

In an effort to secure the participation of more schools and to lessen the burden of participation on the states, ETS and Westat offered to perform all of the work involved, including communicating with school staff, sampling, and administering the assessment.

Twenty-four jurisdictions decided to participate in the field test. Twenty-one of the jurisdictions decided to have NAEP administer all field test sessions. In these jurisdictions, the state coordinator secured the cooperation of the selected schools and then Westat contacted the schools, confirmed the schedule and arrangements, selected the student samples, and conducted the assessment sessions. Three states—Florida, Kentucky, and Wisconsin—chose to have school staff (assessment administrators) conduct the fourth-grade assessments. Westat conducted training session for the assessment administrators in these states. None of the jurisdictions elected to conduct the eighth-grade assessments themselves.

Each participating jurisdiction was asked to appoint a state coordinator to secure the cooperation of sampled schools, and to be the liaison between NAEP/Westat staff and the participating schools.

As described in Chapter 3, the state coordinator for each state was sent the names of approximately 30 pairs of selected schools and requested to secure the cooperation of one school from each pair. This process had been used successfully in the field test for 1990, and was again successful in the field test for 1992. In total, 664 schools agreed to participate in the field test; in 662 of these schools assessment sessions were conducted.

As stated earlier, Florida, Kentucky, and Wisconsin chose to administer the new components of the assessment, fourth-grade reading and mathematics, in order to gain experience with the procedures planned for 1992. The rest of this section describes the procedures implemented in those three states.

Although the three states were responsible for the actual administration at the school level, Westat was responsible for developing the administration materials and procedures and for training state staff. Two training sessions were conducted by Westat home office staff in each of the three states during mid-January. All assessment administrators received a manual before attending one of these training sessions. The training program consisted of a video presentation, scripted lecture and training exercises.

In January 1991, Westat field supervisors selected the student sample for each school and prepared an Administration Schedule (roster) of the sampled students. The Administration Schedule was sent by the state coordinator to the school two weeks before the scheduled assessment date. The other assessment materials were shipped by NCS to arrive two weeks before the scheduled assessment date. Upon receiving the Administration Schedule and the assessment materials, the assessment administrator followed NAEP procedures to select an additional sample of newly enrolled students, identify students who were not capable of participating in the assessment, and prepare assessment questionnaires.

On assessment day, the field supervisor observed the assessment and queried the assessment administrator about the session, procedures, and materials. Supervisors used an Observation Form to record information about the major events related to the assessment and the assessment administrators' opinions and comments.

4.2.2 Results of the Field Test

The overall desired student participation level for the field test was determined from the goal of obtaining 300 student responses for each item to be used in the national assessment and 500 student responses for each item to be used in the Trial State Assessment. Depending on the size of the school, the school's sample numbered approximately 30 to 60 students, who were assessed in either one or two sessions.

Given these goals, the overall desired student participation in both the national and Trial State components of the field test was 22,600 students. In actuality, 24,910 students, or about 10 percent more than required, were assessed.

The field testing of materials and procedures at the fourth-grade level for the Trial State Assessment in the three states provided useful information for NAEP staff in preparation for 1992. While the sessions went well and 80 to 90 percent of assessment administrators thought that the training session, the manuals, and the assessment materials worked well, the administrators did make many suggestions for improving these materials and procedures for the 1992 assessment program.

4.3 THE 1992 TRIAL STATE ASSESSMENT

Forty-one states, the District of Columbia, and two territories volunteered for the 1992 Trial State Assessment. This is a net increase of four jurisdictions over 1990, with seven newly participating in 1992 and three that were in the 1990 assessment deciding not to participate in 1992. Figure 4-1 identifies the jurisdictions participating in each of the two assessment years. As with the field test, each jurisdiction designated a state coordinator to oversee all assessment activities in the state.

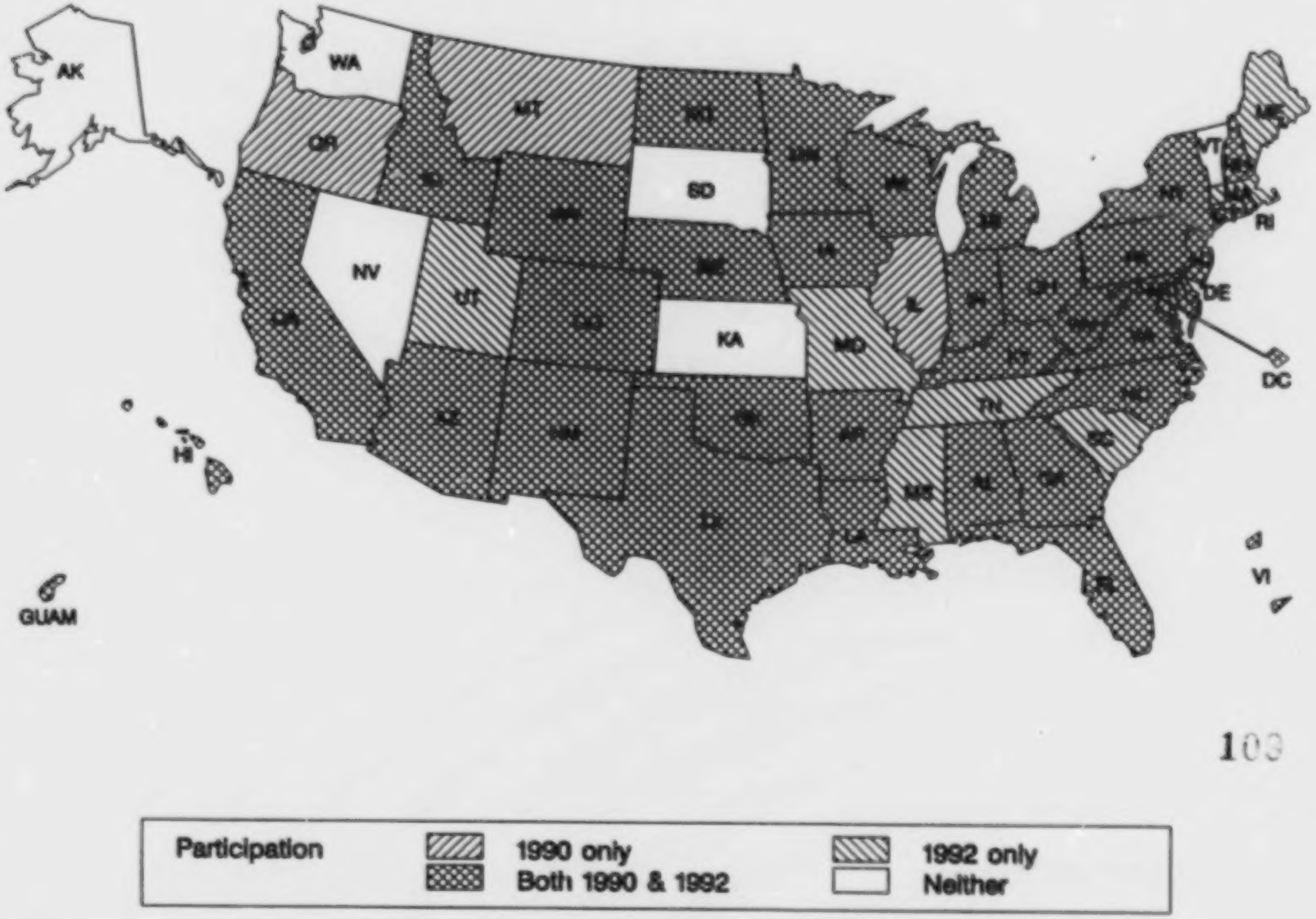
Two states—Illinois and Washington—had agreed to participate in the 1992 Trial State Assessment, but dropped out before the assessment began, primarily due to a lack of success in getting schools in their states to participate. This followed a letter from NCES recommending that states obtain at least a 70 percent school cooperation rate in order to meet the guidelines for participation.

4.3.1 Overview of Responsibilities

The data collection for the 1992 Trial State Assessment involved a collaborative effort between the participating states and the NAEP contractors, especially Westat, the field administration contractor. Westat's responsibilities included

- selecting the sample of schools and students for each participating state;
- developing the administration procedures and manuals;

Figure 4-1
Participating Jurisdictions, 1990 and 1992 Trial State Assessments


AK
WA
OR
MT
SD
NV
UT
CA
KA
MO
IL
TN
SC
MS
ME
VT
RI
DE
DC
HI
GUAM
VI
Participation
1990 only
Both 1990 & 1992
1992 only
Neither

BEST COPY AVAILABLE

- training the state personnel who would conduct the assessments; and
- conducting an extensive quality assurance program.

Each jurisdiction volunteering to participate in the 1992 program was asked to appoint a state coordinator. In general, the state coordinator was the liaison between NAEP/Westat staff and the participating schools. In particular, the state coordinator was asked to

- gain the cooperation of the selected schools;
- assist in the development of the assessment schedule;
- receive the lists of all grade eligible students from the schools;
- coordinate the flow of information between the schools and the NAEP contractors;
- provide space for the state supervisor to use when sampling;
- notify assessment administrators about training and send them their manuals; and
- send the lists of sampled students to the schools.

At the local school level, an assessment administrator was responsible for preparing for and conducting the assessment session(s) in one or more schools. These individuals were usually school or district staff and were trained by Westat staff. The assessment administrator's responsibilities included

- receiving the list of sampled students from the state coordinator;
- identifying sampled students who should be excluded;
- distributing assessment questionnaires to appropriate school staff;
- notifying sampled students and their teachers;
- administering the assessment session;
- completing assessment forms; and
- preparing the assessment materials for shipment.

Westat hired and trained six field managers and 44 state supervisors, one for each jurisdiction. Each field manager was responsible for working with the state coordinators of seven to eight states and for overseeing assessment activities. The primary tasks of the field managers were to

- obtain information about cooperation and scheduling;
- make sure the arrangements for the assessments were set and assessment administrators identified; and
- schedule the assessment administrators training sessions.

The primary tasks of the state supervisors were to

- select the sample of students to be assessed;
- conduct in-person assessment administrator training sessions; and
- coordinate the monitoring of the assessment sessions and makeup sessions.

Westat also hired and trained an average of eight quality control monitors in each state to monitor 50 percent of the assessment sessions.

4.3.2 Schedule of Data Collection Activities

May 15, 1991	Westat sent the samples of schools selected for the National and Trial State Assessment to the state coordinators.
Early August, 1991	Westat field managers visited each state to explain the computerized State Coordinator System, which could be used to keep track of assessment-related activities. Westat distributed Student Listing Forms, Principal Questionnaires, and the list of the schools selected for the Trial State Assessment updated with a suggested week of assessment and number and type of sessions.
May-November, 1991	State coordinators obtained cooperation from districts and schools. State coordinators reported participation status to Westat field managers via printed lists or computer files. State coordinators sent Student Listing Forms, Supplemental Student Listing Forms, and Principal Questionnaires to participating schools.
October-November, 1991	Westat selected substitutes for refusals and sent them to state coordinators. States reporting the participation status of all schools by October 15 received substitutes for refusals by October 31. States reporting by October 31 received substitutes by November 15.

November 14-17, 1991	State supervisor training sessions were held.
December 2-20, 1991	NAEP state supervisors visited state coordinators to select student samples and prepare Administration Schedules listing the students selected for each session. Westat provided schedule of training sessions and copies of the Manual for assessment administrators to state coordinators for distribution.
December 2, 1991-January 10, 1992	State coordinators notified assessment administrators of the date and time of training and sent each a copy of the *Manual for Assessment Administrators*.
January 3-10, 1992	Quality control monitor training sessions were held.
January 9-31, 1992	Assessment administrator training sessions were held.
January 20-February 14, 1992	State coordinators sent Administration Schedules to each school two weeks before the scheduled assessment date.
February 3-28, 1992	Assessments were conducted. Unannounced visits were made by quality control monitors to a predetermined 50 percent of the sessions.
March 2-6, 1992	Makeup sessions were held as necessary.

4.3.3 Preparations for the Trial State Assessment

The focal point of the schedule for the Trial State Assessment was the period between February 3-28, 1992, when the assessments were conducted in the schools. However, as with any undertaking of this magnitude, the project required many months of planning and preparation.

Westat selected the samples of fourth- and eighth-grade schools according to the procedures described in Chapter 3. On May 15, 1991, lists of these selected schools and other materials describing the Trial State Assessment Program were sent to state coordinators. This mailing took place about two months earlier than for the 1990 assessment because state coordinators had requested more time to contact districts and schools and schedule the assessments. Most state coordinators also preferred that NAEP provide a suggested assessment date for each school. School listings were updated with this information and were sent to the state coordinators, along with other descriptive materials and forms, in early August.

State coordinators also were given the option of receiving the school information in the form of a computer database with accompanying management information software. This system enabled the state coordinators to keep track of the cooperating schools, the assessment schedule, the training schedule, and the assessment administrators. Coordinators could choose to receive a laptop computer and printer or to have the system installed on their own computer.

Westat field managers traveled to the state offices to explain the computer system to the state coordinators and their staff. All but one state coordinator chose to receive the system.

Six of the most experienced NAEP supervisors were chosen to be field managers, the primary link between NAEP and the state coordinators. In mid-August, the field managers visited offices of the state coordinators to explain the computerized system to state staff. The field managers kept in frequent contact with the state coordinators as the state coordinators secured the cooperation of the selected schools and established the assessment schedule.

The field managers used the same computer system as the state coordinators to keep track of the schools and schedule. The state coordinators sent updates either via computer disks, by telephone, or in print to their field manager, who then entered in the information into the system. Weekly transmissions were made from the field manager to Westat.

The state coordinators' first task was to secure the participation of the selected schools. States that had determined the cooperation status of all selected schools by October 15 were sent a list of potential replacements for refusals by October 31. States that reported by October 31 received a list of potential substitutes by November 15. Both printed lists and computer files of substitute schools were transmitted to the field managers and state coordinators. (See Chapter 3 for more details about school substitution.)

In mid-November, Westat hired one state supervisor for each participating state. The state supervisors attended a training session held in the Washington, DC, area between November 14-17, 1991. This training session focused on the state supervisors' immediate tasks—selecting the student samples and hiring quality control monitors. State supervisors also were given the training script and materials for the assessment administrators' training sessions they would conduct in January so they could begin to become familiar with these materials.

The state supervisors' first task after training was to complete the selection of the sample of students who were to be assessed in each school. All participating schools were asked to send a list of their grade-eligible students to the state coordinator by November 15. Sample selection activities were conducted in the state coordinator's office unless the state coordinator preferred that the lists be taken to another location.

Using a preprogrammed calculator, the supervisors generally selected a sample of 30 students per session type per school. The exceptions to this were small schools and states with fewer than the necessary 100 eighth-grade or 125 fourth-grade schools. In the states with fewer schools, larger student samples were required from schools that participated.

After the sample was selected, the supervisor completed an Administration Schedule for each session, listing the students to be assessed. The Administration Schedules for each school were put into an envelope and given to the state coordinator to send to the school two weeks before the schedule assessment date. Included in the envelope were instructions for sampling students who had enrolled at the schools since the creation of the original list used in sampling.

During the period from mid-November through December, the state supervisors also recruited and hired quality control monitors to work in their states. It was the quality control monitor's job to observe the sessions designated to be monitored, complete an observation form on each session and to intervene when the correct procedures were not followed. In each state, half of the sessions were designated to be monitored. This information was known only to contractor staff; it was not on any of the listings provided to state staff.

Approximately 400 quality control monitors were trained in two training sessions held during January 3-7 and 7-10, 1992. The first day of each training session was devoted to a presentation of the assessment administrators training program by the state supervisors, which not only gave the quality control monitors an understanding of what assessment administrators were expected to do, but gave state supervisors an opportunity to practice presenting the training program. The remaining days of the training sessions were spent reviewing the quality control monitor observation form and the role and responsibilities of the quality control monitors.

Almost immediately after the quality control monitor training sessions, the supervisors began conducting the assessment administrator training sessions. Each quality control monitor attended several of these training sessions, to assist the state supervisor and to become thoroughly familiar with the assessment administrator's responsibilities. Almost 10,000 persons who were to be assessment administrators were trained in about 500 training sessions across the nation.

To ensure uniformity in the training sessions, Westat developed a highly structured program involving a script for trainers, a videotape, and a training example to be completed by the trainees. The supervisors were instructed to read the script verbatim as they proceeded through the training, ensuring that each trainee received the same information. The script was supplemented by the use of overhead transparencies, displaying the various forms that were to be used and enabling the trainer to demonstrate how they were to be filled out.

The videotape, similar to the one used in the 1990 Trial State Assessment, was developed by Westat to provide background for the study and to simulate the various steps of the assessment that would be repeated by the assessment administrators. The portions of the videotape depicting the actual assessment had been taped in a classroom with students in attendance to closely simulate an actual assessment session. The videotape was divided into sections with breaks for review by the trainer and practice for the trainees.

The final component of the presentation was the "Training Example." This consisted of a set of exercises keyed to each part of the training package. A portion of the videotape was shown and then reviewed by the trainer following the script. Then, exercises related to that material were completed by the trainees before the next subject was discussed.

The entire training session generally ran for about three and one-half hours. Sessions usually began in the morning and ended with lunch. In 1990, the training sessions had generally lasted about five to six hours. Responding to requests from state coordinators and assessment administrators, Westat trimmed the training session to one-half day.

All of the information presented in the training session was included in the *Manual for Assessment Administrators*, developed by Westat. There were two versions of the manual, one for each grade. Copies of the manuals were sent by Westat to the state coordinators at the beginning of December so that they could be distributed to the assessment administrators before the training sessions.

4.3.4 Monitoring of Assessment Activities

Two weeks prior to the scheduled assessment date, the assessment administrator received the Administration Schedule and assessment questionnaires and materials. Five days before the assessment, the quality control monitor made a call to the assessment administrator and recorded the results of the call on the Observation Form. Most of the questions asked in the pre-assessment call were designed to gauge whether the assessment administrator had received all materials needed and was prepared for the session.

Pre-assessment calls were made to all schools regardless of whether they were to be monitored. If the sessions in a school were not observed, the quality control monitor called the assessment administrator three days after the assessment to find out how the session went, to obtain the assessment administrator's impressions of the manual, training, and materials and to ensure that all post-assessment activities had been completed.

If the sessions in a school were to be monitored, the quality control monitor was to arrive at the school one hour before the scheduled beginning of the assessment to observe preparations for the assessment. To ensure the confidentiality of the assessment items, the booklets were packaged in shrink-wrapped bundles and were not to be opened until the quality control monitor arrived or 45 minutes before the session began, whichever occurred first.

In addition to observing the opening of the bundles, the quality control monitor used the Observation Form to check that the following had been done correctly: sampling newly enrolled students, reading the script, distributing and collecting assessment materials, timing the booklet sections, answering questions from students, and preparing assessment materials for shipment.

After the assessment was over, the quality control monitor obtained the assessment administrator's opinions of how the session went and how well the materials and forms worked.

If four or more students were absent from the session, a makeup session was to be held. If the original session had been monitored, the makeup session was also monitored. This required coordination of scheduling between the quality control monitor and assessment administrator.

4.3.5 School and Student Participation

Table 4-1 shows the results of the state coordinators' efforts to gain the cooperation of the selected schools. Overall, almost 9,000 schools—4,921 for grade 4 and 3,798 for grade 8—participated in the 1992 Trial State Assessment. This is about 88 percent (unweighted) of the eligible schools in the original sample at each grade and about 95 percent (unweighted) of the sample after substitution.

Table 4-1

School Participation, 1992 Trial State Assessment

Status	Grade 4	Grade 8
Schools in original sample	5356	4118
Schools not eligible (e.g., closed, no grade 4/8)	152	128
Eligible schools in original sample	5204	3990
Noncooperating (e.g., school, district, state refusal)	605	471
Participating	4599	3519
Substitutes provided for noncooperating schools	501	409
Participating substitutes	322	279
Total schools participating after substitution	4921	3798

Table 4-2

Student Participation in the 1992 Trial State Assessment of Mathematics

Status	Grade 4 Mathematics	Grade 8 Mathematics
Sampled	128,770	129,239
Original sample	125,008	125,725
Supplemental sample	3,762	3,514
Withdrawn	5,545	6,087
Excluded	6,424	6,532
To be assessed	116,801	116,620
Assessed	111,276	108,557
Initial sessions	110,970	107,461
Makeup sessions	306	1,096

Participation results for students in the 1992 Trial State Assessment are given in Table 4-2. Approximately 129,000 students were sampled in each grade. As can be seen from the table, the original sample, which was selected by the NAEP state supervisors, comprised about 125,000 of this number. The original sample size was increased somewhat after the supplemental samples had been drawn (from students newly enrolled since the creation of the original lists).

Assessment administrators removed some students from the total sample according to NAEP criteria: first, those students who had left their schools since the time that they were sampled (withdrawn); then, those judged incapable of participating meaningfully in the assessment by school staff (excluded). A student could be excluded if she or he either had an Individualized Education Plan (IEP) or was classified as Limited English Proficient (LEP), was incapable of participating meaningfully, and met certain other criteria.

These exclusions left 116,801 fourth graders and 116,620 eighth graders to be assessed in mathematics. Of these, 111,276 fourth graders and 108,557 eighth graders were assessed, yielding unweighted student participation rates of 95.3 percent and 93.1 percent, respectively.

4.3.6 Results of the Observations

During the assessment sessions, the quality control monitors were to note instances when the assessment administrators deviated from the prescribed procedures and whether any of these deviations were serious enough to warrant their intervention. Quality control monitors reported no instances where there were serious breaches of the procedures or major problems that would question the validity of the assessment.

Prescribed procedures were most often deviated from in the administrator's reading of the script that introduced the assessment and provided the directions. Even so, in at least 90 percent of the observed sessions the assessment administrator read the script verbatim or with only slight deviations. Examples of major deviations included skipping sections of the script, adding substantially to the script, and forgetting to pass out materials at the appropriate times. The quality control monitor intervened in these instances.

Most of the other procedures that could have had some bearing on the validity of the results were adhered to very well by the assessment administrators. In 99 percent of the observed sessions, the assessment administrators opened the bundles of booklets at the appropriate time and handled questions from the students correctly. Ninety-nine percent of the fourth-grade and 98 percent of the eighth-grade sessions were timed correctly.

In 95 percent of the observed mathematics sessions at both grades, the assessment administrator handled the distribution and collection of calculators without problems. In 95 percent of the fourth-grade mathematics sessions and 97 percent of the eighth-grade mathematics sessions, the assessment administrator conducted the calculator training without problems.

After the assessment session was over, assessment administrators were asked how they thought the assessment went and whether they had any comments or suggestions. Overall, assessment administrators stated that they thought 98 to 99 percent of the sessions went very well or satisfactorily.

Assessment administrators reported that fewer of the fourth-grade mathematics sessions (74%) went very well compared with the eighth-grade sessions (86%). The percent of monitored sessions versus unmonitored sessions that assessment administrators thought went very well was slightly higher at the fourth grade (78% compared to 70%), but remained the same (86%) for the eighth grade.

Comments about the assessment materials and procedures were generally favorable. Criticisms or suggestions included that there were too many forms and too much paperwork; coding the booklet covers was tedious and problematic for students; and schools needed more information about NAEP and assessment results.

In addition to these interviews, Westat sent a debriefing form to all of the NAEP state supervisors and met in person with half of them. This meeting produced suggestions for future assessments, especially many minor changes in the procedures, materials and training plans. In addition, the state supervisors recommended that district and particularly school staff receive more information describing the background and objectives of NAEP and the Trial State Assessments. They also stated that many school staff were very interested in results for their students, or at least summary results for their state.

State coordinators were also sent a questionnaire about their experiences, suggestions, and comments. State coordinators from 39 of the participating states and territories responded. All of the 35 state coordinators responding to the question "How did the assessments go in your state?" said "Very well" to "Fairly well." They also commented favorably on the training package and other materials. Like the assessment administrators, the state coordinators criticized the amount of work required to prepare for the assessments. They made many other suggestions about the computerized data system, sampling procedures, training program, and design of the assessment. All of these suggestions will be reviewed as future assessments are planned.

The results of the assessment and comments from assessment administrators and state coordinators were summarized in a report presented to the NAEP Network on May 11, 1992. In mid-August, each participating state and territory received a summary of its participation data, data collection activities, results of the assessment, and assessment administrators' comments.

Chapter 5

PROCESSING ASSESSMENT MATERIALS

Dianne Smrdel, Linda Reynolds, and Brad Thayer

National Computer Systems

5.1 OVERVIEW

This chapter describes the printing, distribution, receipt, processing and final disposition of materials for the mathematics portion of the Trial State Assessment. The scope of the effort required by National Computer Systems (NCS) to process the materials is evidenced by the following:

- Prior to the assessment, 13,448 bundles of assessment booklets at grade 4 and 13,070 bundles at grade 8 were created and distributed to approximately 9,000 schools.
- For the approximately 111,000 students assessed for grade 4, about 222,000 assessment booklets and 35,800 questionnaires were received and processed; and about 2,000,000 student responses from 59 constructed-response items were professionally scored.
- For the approximately 109,000 students assessed for grade 8, about 218,000 assessment booklets and 22,500 questionnaires were received and processed; and about 2,050,000 student responses from 65 constructed-response items were professionally scored.
- In all, approximately 7 million double-sided pages from test booklets and questionnaires were optically scanned.

Throughout the processing, the NCS Process Control System and Workflow Management System were used to track, audit, edit, and resolve characters of information. A quality control sample of characters of transcribed data was selected and compared to the actual responses in the assessment booklets.

The volume of collected data and the complexity of the Trial State Assessment processing design, with its spiraled distribution of booklets, as well as the concurrent administration of this assessment and the national assessments, required the enhancement and implementation of flexible, innovatively designed processing programs and a sophisticated Process Control System. This system, developed for the 1990 assessments, allowed an integration of data entry and workflow management systems, including carefully planned and delineated editing, quality control, and auditing procedures.

The magnitude of the effort is apparent when considering that the activities described in this chapter were completed concurrently with the processing of the national assessments, that all processing activities were completed within 10 weeks, and that an estimated accuracy rate of fewer than five errors for every 10,000 characters of information was achieved.

Several major changes in materials processing were made from 1990, including the conversion of all documents to scannable form, the tailoring of shipments to the individual size and requirements of schools, and the reorganization of the process flow to conduct constructed-response scoring after all machine scoring and data verification processes were complete, allowing NCS to provide Westat and ETS with demographic and cognitive data at an earlier date.

5.2 PROCESS CONTROL SYSTEM

NCS maintains a Process Control System consisting of numerous specialized programs and processes to accommodate the unique demands of concurrent assessment processing and a unified ETS/NCS system integration. The Process Control System, which was developed for the 1990 assessment, was necessary to maintaining control of all shipments of materials to the field, of all receipt from the field, and of any work in progress. The system is a unique combination of several reporting systems currently in use at NCS, along with some application-specific processes. These systems are the Workflow Management System, the Bundle Assembly Quality Control System, the Outbound Mail Management System, and the On-line Inventory Control system. Data were collected from these systems and recorded in the file called the "NAEP Process Control System." Additional information was directly entered into the Process Control System.

5.3 WORKFLOW MANAGEMENT SYSTEM

The functions of the Workflow Management System are to keep track of where the production work is and where it should be and to collect data for status reporting, forecasting, and other ancillary subsystems. The primary purpose of the Workflow Management System is used to analyze the current workload by project across all work stations.

The data processing and control systems are determined to a large extent by the type of documents processed. For the Trial State Assessment, only machine-scannable assessment booklets and answer documents were used to collect student responses. The five questionnaires that were used to collect data about school characteristics, teachers associated with sampled students, and students excluded from the assessment were also scannable documents.

5.4 PROCESS FLOW OF NAEP MATERIALS AND DATABASE CREATION

Figure 5-1 shows the conceptual framework of processes that were used both for the Trial State Assessment materials and for the national NAEP materials.

Section I of Figure 5-1 depicts the flow of NAEP's printed materials. Information from the Administration Schedule and Packing List was used to control the processing of materials. The figure follows the path of each assessment instrument—Student Test Booklets, School Characteristics and Policies Questionnaires, Teacher Questionnaires, Excluded Student Questionnaires, Packing List, and Administration Schedules—as they were tracked through the appropriate processes that resulted in the final integrated NAEP database.

The remainder of this chapter provides an overview of the materials processing activities as shown in Section I of Figure 5-1 and detailed in Figure 5-2. Section II of Figure 5-1 depicts the evolution of the NAEP/NCS database from the transcribed data to the final files, provided to Westat for creation of weights and to ETS for analysis and reporting.

The 1992 NAEP data collection resulted in six classes of data files (student, school, teacher, excluded student, sampling weight, and item information files). The structure and internal data format of the 1992 NAEP database was a continuation of the integrated design originally developed by ETS in 1983.

5.5 MATERIALS DISTRIBUTION

The use of bar code technology in document control was introduced to NAEP by NCS in the 1990 assessment; its use continued in 1992. Bar codes were applied to the front cover of the documents. The bar code consisted of the two-digit booklet number, a five-digit sequential number, and a check digit. It was unnecessary to pre-identify the estimation booklets with bar codes because students were instructed to grid the estimation booklet cover with the identification number of their original booklet.

The booklets were spiraled into 26 unique bundles consisting of 11 booklets in a set pattern. A header sheet was attached to each bundle that indicated the assessment type, bundle type, bundle number, and a list of the booklet types to be included in the bundle.

The bundle numbers on the header sheet were created to identify the type of bundle. All bundles were then passed under a scanner programmed to interpret this type of bar code and the file of scanned barcodes was transferred from the scanner to the mainframe. A computer program compared the bundle type expected to the one actually scanned after the header and verified that there were 11 booklets in each bundle. Any discrepancies were printed on an error listing forwarded to the Packaging Department, where the error was corrected and the bundle was again read into the system for another quality control check. This process was repeated until all bundles were correct.

The bundles were shrink-wrapped in clear plastic. The estimation booklets were also shrink-wrapped in groups of 11. A bundle of pre-identified booklets and a bundle of estimation booklets were strapped together. A bright label was placed over the cross of the straps that

Figure 5-1

Data Flow Overview, 1992 Trial State Assessment

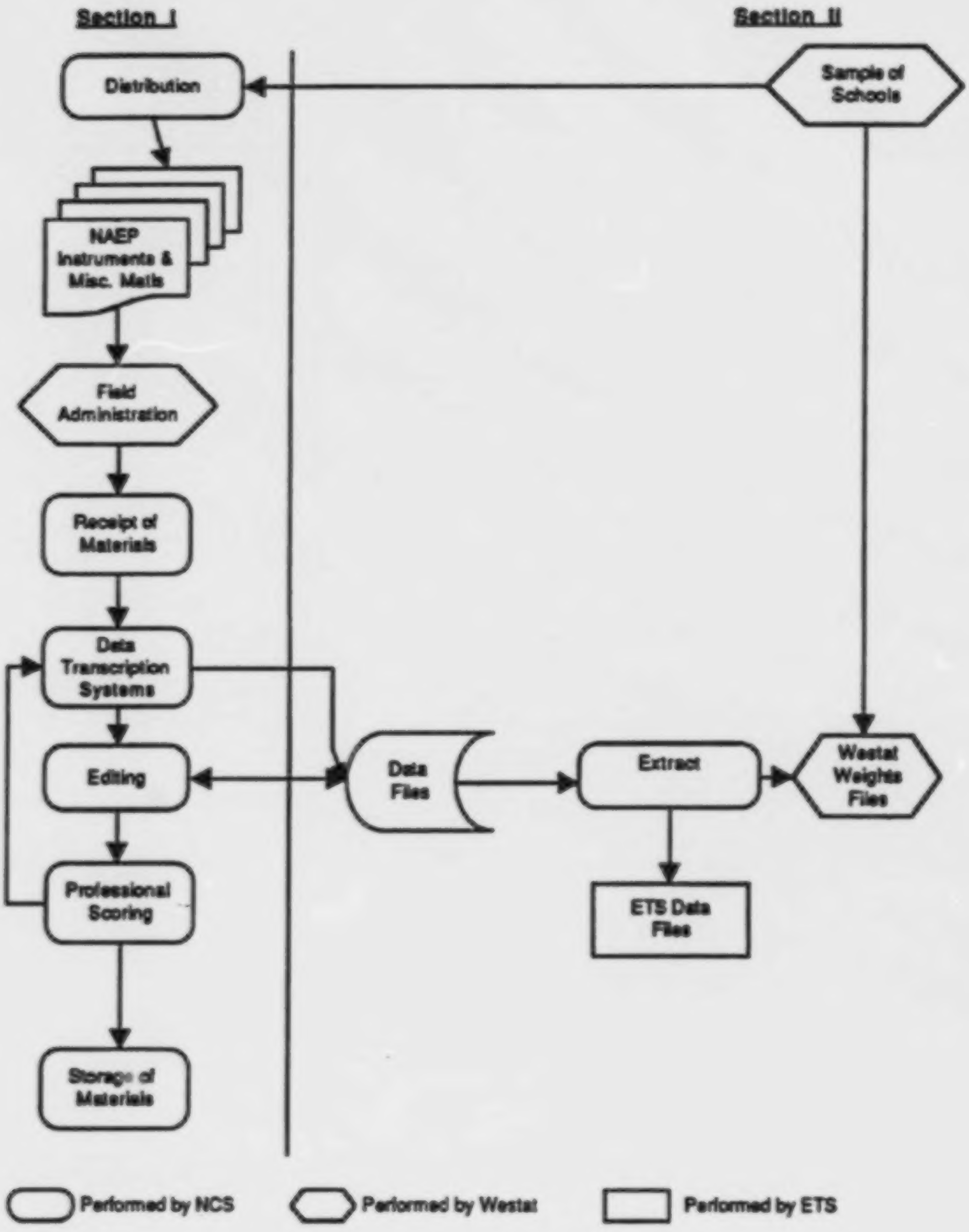
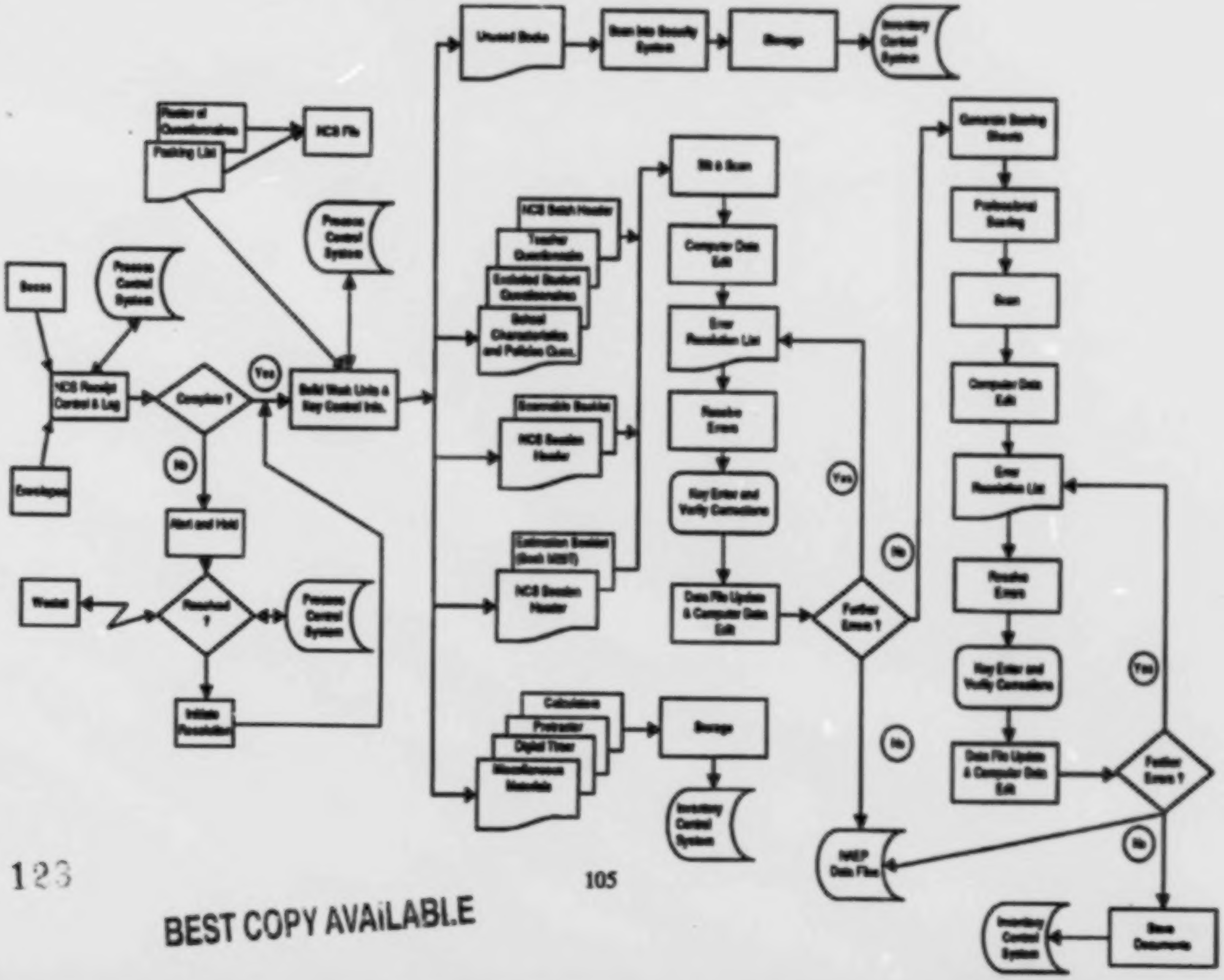

Section I
Section II
Distribution
Sample of Schools
NAEP Instruments & Misc. Matls
Field Administration
Receipt of Materials
Data Transcription Systems
Editing
Professional Scoring
Storage of Materials
Data Files
Extract
Westat Weights Files
ETS Data Files
Performed by NCS
Performed by Westat
Performed by ETS

Figure 5-2

Materials Processing Flow, 1992 Trial State Assessment


Unused Books
Scan Into Security System
Storage
Inventory Control System
Roster of Questionnaires
Packing List
NCS File
Process Control System
Boxes
Process Control System
NCS Receipt Control & Log
Complete ?
Yes
Build Work Units & Key Control Info.
Envelopes
No
Alert and Hold
Waste
Resolved ?
Process Control System
Initiate Resolution
NCS Batch Header
Teacher Questionnaire
Excluded Student Questionnaires
School Characteristics and Policies Ques.
Scannable Booklet
NCS Session Header
Estimation Booklet
NCS Session Header
Calculators
Protractor
Digital Timer
Miscellaneous Materials
Storage
Inventory Control System
Slit & Scan
Computer Data Edit
Error Resolution List
Resolve Errors
Key Enter and Verify Corrections
Data File Update & Computer Data Edit
Further Errors ?
Yes
No
No
NAEP Data Files
Generate Scoring Sheets
Professional Scoring
Scan
Computer Data Edit
Error Resolution List
Resolve Errors
Key Enter and Verify Corrections
Data File Update & Computer Data Edit
Further Errors ?
Yes
No
Store Documents
Inventory Control System

BEST COPY AVAILABLE

read "Do Not Open Until 45 Minutes Before Testing." Following this, bundles were ready for assignment and distribution.

When packing lists for distribution of materials were created from the Materials Distribution System, a second and more detailed bundle slip was produced. This bundle slip indicated the same information as the slip wrapped with the bundle, in addition to the school number and the complete booklet ID numbers of the booklets within that bundle. This allowed the assessment administrators to pre-assign booklets for their sessions.

The timing of the shipments of these materials to the participating schools was critical, since the shipments needed to be in the school at least one week but not more than two weeks prior to testing. As in 1990, calculators were in limited supply. Therefore, shipments for assessments occurring during the last week could not be completed until shipments from the first week's assessments were returned. This affected the shipments for both grade 4 and grade 8.

Each school conducted at least one session; some conducted more than one. The materials needed for a school to conduct all of its mathematics sessions were sent in one shipment. The booklets for the fourth-grade reading session(s) were boxed separately in the same shipment. In 1990, each session's materials had been shipped independently. Although this change in shipment practice eliminated the option to pre-assemble many materials, it did cause less confusion within the schools.

Some materials were distributed per school; others were distributed per session. Materials issued per session were:

	Bundle(s) of 11 assessment booklets (based on sample count)
6	Scientific calculators (grade 8) per bundle of booklets
6	Simple calculators (grade 4) per bundle of booklets
15	Protractors
1	Cassette tape for estimation booklet
1	Digital timer
1	Calculator poster
1	Mathematics poster
1	Tape recorder with batteries
5	Rulers
5	Sets of geometric shapes
1	Pad of appointment cards
1	Return postage paid label
1	Post-it note pad
1	Shipping tape
5	Excluded Student Questionnaires
5	Teacher Questionnaires

Those materials distributed by school were:

2	Roster of Questionnaires
2	Assessment Notifications
1	Pre-addressed envelope
1	School Characteristics and Policies Questionnaire
1	Pre-addressed box

Shipments were sent according to the week of assessment. Some schools found they needed extra quantities of materials (i.e., more excluded student questionnaires or more teacher questionnaires) and calls were received requesting these additional materials.

Aiding in the security of the shipments was the decision to send all shipments, whenever possible, through Airborne. NCS is connected to the Airborne system through computer link thus expediting tracing of any misdirected shipments. This system provides the date and time of delivery as well as the name of the person who signed for the shipment. All shipments were recorded in the Airborne Libra system. If a shipment had to be sent by UPS or the U.S. Postal Service, this information was also recorded and transferred to the mainframe.

5.6 PROCESSING ASSESSMENT MATERIAL

The materials from each session were to be returned to NCS in the same box in which they were originally mailed. It was the responsibility of the assessment administrator in the unmonitored schools and the quality control monitor in the monitored schools to repackage the items in the proper order, complete all paperwork and return the shipment through the U.S. Postal Service, using the postage-paid label provided.

With approximately 9,000 individual shipments arriving over a four-week period, it was necessary to devise a system that would quickly acknowledge receipt of a school's material. A label applied to the outside of the box by the NCS packaging department contained a bar code which indicated the school number and the project number. When the shipment arrived at NCS, the bar code was read and the shipment forwarded to the receiving area. The file was then transferred to the mainframe through a PC link and a computer program was used to apply the shipment receipt date to the appropriate school within the Process Control System. This provided current status of shipments received regardless of any processing backlog. This information was then transferred electronically to Westat. The status of the administration was checked and in some cases a trace was initiated on the shipment.

Receiving personnel also checked the shipment to verify that the contents of the box matched the school and session indicated on the label. Each shipment was checked for completeness and accuracy, regardless of whether it was monitored or unmonitored.

The materials were checked against the Packing List (see Figure 5-3) to verify that all materials were returned. If any discrepancies were found, an alert was issued. If all assessment instruments were returned, processing continued. Quantities of scientific calculators were in short supply; therefore, during the first two weeks of the assessment, calculators were taken from the incoming shipments and returned to the packaging area to be included in other shipments for the last weeks of testing.

Each booklet and Excluded Student Questionnaire was verified against the Administration Schedule. This included verification of all counts of booklets returned and the matching of information on the front cover of the booklets to that on the Administration Schedule. If any discrepancy was discovered, an alert was issued. The same verification was followed to assure that one estimation booklet was received for every student assessed, and that the correct booklet number was gridded on the front cover to assure matching.

After the contents of the shipment had been identified and verified, the information from the Administration Schedule was entered into the Process Control System. That information included school number, session code, counts of the number of students in original sample, supplemental sample, total sample, withdrawn, excluded, to be assessed, absent, original assessed, assessed in makeup and total assessed. If a makeup session was expected, an information alert was issued to facilitate tracking. The control counts were used by NCS for verification of processing counts. This information was also transferred electronically to Westat on a weekly basis to be used to produce participation statistics for the states.

If quantities and individual information matched, the booklets were organized into work units and batched for processing. The processing flow was changed in 1992, resulting in the completion of the machine scoring prior to the constructed-response scoring. Each batch, consisting of multiple sessions, was assigned a unique batch number. The batch number was entered on the Workflow Management System, facilitating the internal tracking of the session and allowing departmental resource planning. A scannable session header, included in the shipment from the school, was coded with the session code and placed on top of the stack of documents. All student documents were forwarded to machine scanning functions. Control documents were forwarded to appropriate record filing systems.

The estimation block was administered in a separate booklet (Book M29T). The cover of Book M29T contained a section in which students recorded the booklet ID of their assigned assessment booklet. The openers verified that the corresponding booklet ID was correctly recorded on the cover and that an estimation booklet had been issued to each assessed student. As Book M29T contained no constructed responses, they were batched separately and forwarded to the scanning area.

The excluded student questionnaires and teacher questionnaires were compared to the Roster of Questionnaires and the Administration Schedule to verify demographic information. Some questionnaires may not have been available for return with the shipment. These were returned to NCS at a later date in an envelope provided for that purpose. If the Excluded Student Questionnaire was not returned with the shipment of booklets, a record containing all demographic information on that student from the Administration Schedule was entered into the Process Control System. If the questionnaire was subsequently returned, this record was deleted. Otherwise, the record was provided to Westat for use in the weighting process.

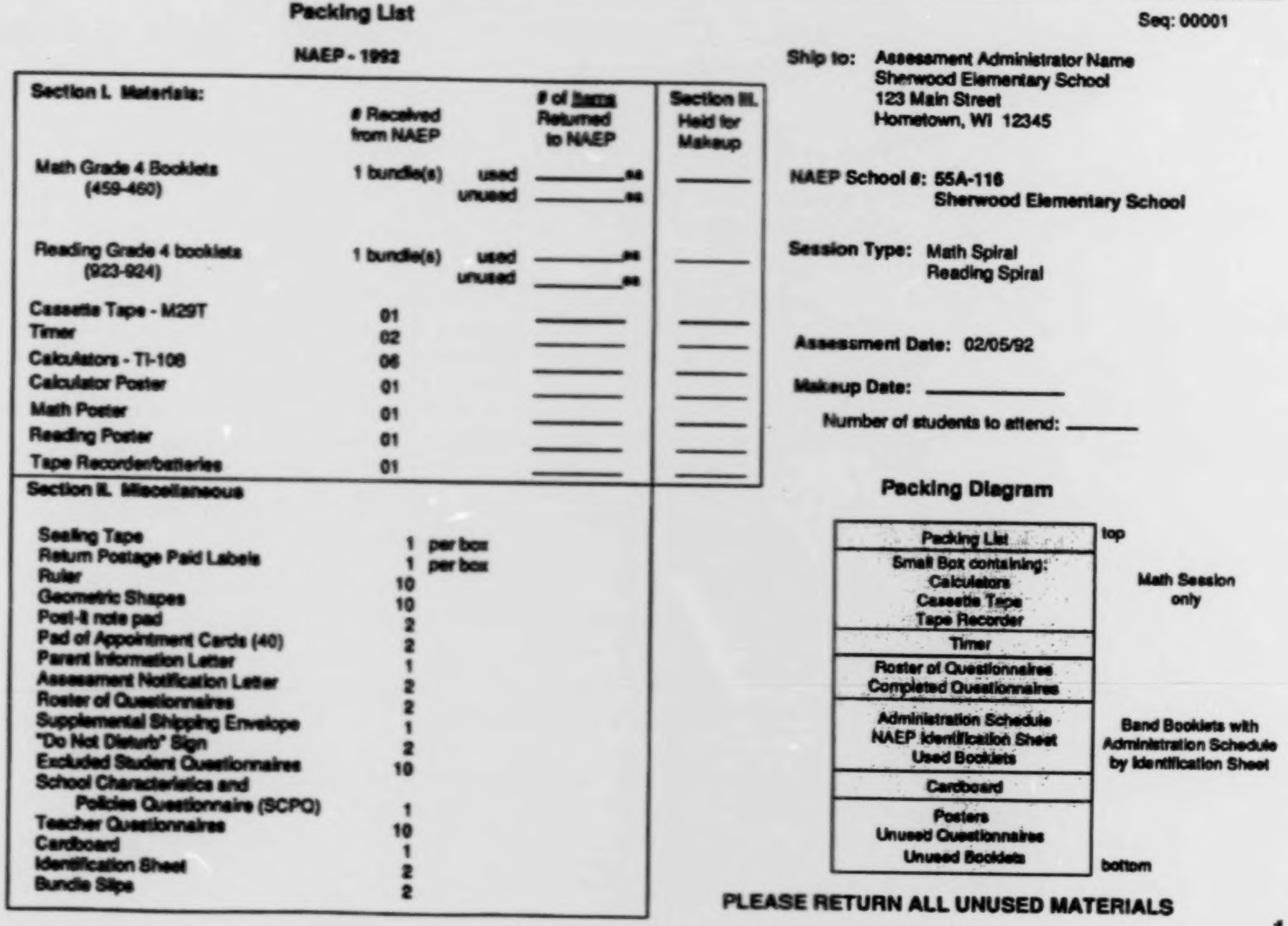
Each school characteristics and policies questionnaire was compared with the Roster of Questionnaires and the school number was verified to match all other materials in the shipment. As with the other questionnaires, this document may not have been returned with the shipment and could also be returned in the supplemental envelope. There was no additional effort made to collect or report information on unreturned school questionnaires.

All assessed and absent students were assigned a test booklet. To indicate an absence, the "A" bubble in the Administration Code column on the front cover of the booklet was gridded. The booklet was then processed with assessed student booklets to maintain session integrity.

The Packing List (Figure 5-3) was used by the schools to account for all materials received from and returned to NCS. Any discrepancies in quantities received or returned to NCS were indicated. Also indicated was whether a makeup session was to be held, the date of scheduled makeup, the number of students involved, and the quantities of materials being held for later return.

Figure 5-3

Packing List, 1992 Trial State Assessment



Packing List

NAEP - 1992

Seq: 00001

Ship to: Assessment Administrator Name
Sherwood Elementary School
123 Main Street
Hometown, WI 12345

NAEP School #: 55A-116
Sherwood Elementary School

Session Type: Math Spiral
Reading Spiral

Assessment Date: 02/05/92

Makeup Date: ____________

Number of students to attend: ______

Section I. Materials:	# Received from NAEP		# of Items Returned to NAEP	Section III. Held for Makeup
Math Grade 4 Booklets (459-460)	1 bundle(s)	used	________ea	________
		unused	________ea	
Reading Grade 4 booklets (923-924)	1 bundle(s)	used	________ea	________
		unused	________ea	
Cassette Tape - M29T	01		________	________
Timer	02		________	________
Calculators - TI-108	06		________	________
Calculator Poster	01		________	________
Math Poster	01		________	________
Reading Poster	01		________	________
Tape Recorder/batteries	01		________	________

Section II. Miscellaneous		
Sealing Tape	1	per box
Return Postage Paid Labels	1	per box
Ruler	10	
Geometric Shapes	10	
Post-it note pad	2	
Pad of Appointment Cards (40)	2	
Parent Information Letter	1	
Assessment Notification Letter	2	
Roster of Questionnaires	2	
Supplemental Shipping Envelope	1	
"Do Not Disturb" Sign	2	
Excluded Student Questionnaires	10	
School Characteristics and Policies Questionnaire (SCPQ)	1	
Teacher Questionnaires	10	
Cardboard	1	
Identification Sheet	2	
Bundle Slips	2	

Packing Diagram

Packing List	top
Small Box containing: Calculators Cassette Tape Tape Recorder	Math Session only
Timer	
Roster of Questionnaires Completed Questionnaires	
Administration Schedule NAEP Identification Sheet Used Booklets	Band Booklets with Administration Schedule by Identification Sheet
Cardboard	
Posters Unused Questionnaires Unused Booklets	bottom

PLEASE RETURN ALL UNUSED MATERIALS

The Administration Schedule contained the demographic characteristics of the students selected for the assessment. This information included the sex, race/ethnicity, birth date, and IEP/LEP indicators. The booklet number of the student selected was recorded on the Administration Schedule during the assessment process, and the demographic information was transferred to the booklet covers by either the student or the assessment administrator.

The demographics of the sampled students who did not participate in the assessment (exclusions and absentees) were provided to Westat to be used to adjust the sampling weights of the students who did participate. The excluded student information was obtained from the excluded student questionnaire or provided on a file for those not returned to NCS. The absent student information was taken from the front cover of the booklet that was assigned prior to the start of the assessment. This procedure eliminated the need for an additional form for absent students.

For the Rosters of Questionnaires, two numbers were entered for each type of questionnaire: number of questionnaires expected and number actually received. The Packing List, Administration Schedule, and Roster of Questionnaires were forwarded to the operations coordinator and filed by school within state for future reference.

5.7 PROFESSIONAL SCORING

The 1992 Trial State Assessment in mathematics contained three different types of cognitive items: extended constructed-response, short constructed-response, and multiple-choice. These items were administered in scannable assessment booklets that were identical to those used for the fourth- and eighth-grade national assessments.

Scores for the constructed-response items were gridded by the readers on separate, scannable scoring sheets, one sheet per booklet. As batches of test booklets cleared the editing process, scoring sheets for each batch of booklets were automatically generated by the system. Since the system had already captured all scannable information from each test booklet, scoring sheets could be generated for only those student booklets for which the student was present and eligible for the assessment. At the same time that the full set of scoring sheets were generated, a 20 percent (minimum) subset of booklets were selected at random by the system for reliability scoring. A separate set of scoring sheets was generated for these booklets.

Once a batch of scoring sheets was matched with the corresponding batch of student assessment booklets, the booklets were forwarded to the professional scoring area. The scoring of the Trial State Assessment was conducted simultaneously with the scoring of the mathematics portion of the national program. The same readers scored the constructed-response items from both programs.

5.7.1 Description of Scoring

Each constructed-response item had a unique scoring guide that identified the range of possible scores for the item and defined the criteria to be used in evaluating the students' responses. Team leaders reviewed discrepancies between readers and reviewed decisions regularly so that all readers scored each item similarly.

The readers scoring the short constructed-response items were organized into eight teams, each comprising 11 readers and a team leader. These teams scored responses to 84 discrete short constructed-response items at the fourth and eighth grades. Of these items, 27 were categorized as right/wrong, while the remaining 57 items included several different categories of correct and incorrect responses. For the items scored as right/wrong, a correct response was scored as 8 and an incorrect response was scored as 1. Items with two correct responses were given a score of 7 for the second correct response. Various types of incorrect responses were also tracked with separate score points. The incorrect responses were assigned a score point from 1 to 5 to capture information on the specific types of errors students were making.

The readers scoring the extended constructed-response items were organized into three teams. One team comprised 11 readers and one team leader. This team scored responses to the 11 discrete extended constructed-response items from the fourth and eighth grades for both the national and the Trial State assessments. The other two teams comprised six readers and one team leader each. These two teams scored responses to 17 discrete extended constructed-response items from the fourth, eighth, and twelfth grades. (Figure 5-4 shows the text of an eighth-grade extended constructed-response item and its scoring guide.) The extended constructed-response items were scored on a rising scale of 1 to 4. Responses that were "off-task" or completely incorrect received a score of 9. In this way, information was captured about what parts of an item students were or were not able to complete correctly.

5.7.2 Training

The readers were trained to ensure that they would reliably score the constructed-response items. The training, which was conducted during a one-week period, familiarized the group with the scoring guides in order to reach a high level of agreement among the readers.

Before the training program began, the team leaders worked with ETS mathematics test development staff to prepare training sets (sets of sample responses to accompany the scoring guides). Training involved explaining each item and its scoring guide to the readers and discussing responses that were representative of the various score points in the guide. The training was conducted by ETS mathematics test development specialists with assistance from the team leaders. Following the explanations, the readers scored and discussed 5 to 35 carefully selected "practice papers" for each item, depending on the complexity of the item. Next, each reader practiced by scoring all the constructed-response items in each of approximately 12 bundles of booklets, with an average of 54 booklets per bundle. (It was not necessary to use this method of training for the items with straight numeric answers since determining the correctness of the student responses for these items was straightforward and fully explained within the scoring guide.) During this practice, discussion sessions were held to review responses that received a wide range of scores.

Figure 5-4
Sample Extended Constructed-response Item and Scoring Guide

This question requires you to show your work and explain your reasoning. You may use drawings, words, and numbers in your explanation. Your answer should be clear enough so that another person could read it and understand your thinking. It is important that you show all your work.

Treena won a 7-day scholarship worth $1,000 to the Pro Shot Basketball Camp. Round trip travel expenses to the camp are $335 by air or $125 by train. At the camp she must choose between a week of individual instruction at $60 per day or a week of group instruction at $40 per day. Treena's food and other expenses are fixed at $45 per day. If she does not plan to spend any money other than the scholarship, what are all choices of travel and instruction plans that she could afford to make? Explain your reasoning.

Solution:

Treena's fixed expenses will be 7 x 45 = $315 for the 7 days. Therefore, she has 1000 - 315 = 685 to spend for instruction and travel. The group plan will cost 7 x 40 = 280 while the individual plan will cost 7 x 60 = 420. Treena has three options:

Group and Train: 280 + 125 = 405 (720)—$280 left
Group and Plane: 280 + 335 = 615 (930)—$70 left
Individual and Train: 420 + 125 = 545 (860)—140 left

She cannot choose the individual plan and travel by plane because her total expenses would be $1,070 which is greater than the allotted scholarship. (This can be considered as a valid conclusion but can only be counted in a score of 1 or 2.) Any full credit response clearly communicates that Treena has three options, what the three options are, and how the student arrived at the three options.

1 a) Student indicates valid conclusions with no mathematical evidence
OR
starts some correct mathematics beyond computing fixed cost (7 x 45 = 315) but indicates no conclusion.

b) Student work contains major mathematical errors or flaws in reasoning. For example: The student does not consider Treena's fixed expenses or does not realize that 40 and 60 must each be multiplied by 7.

2 a) Student indicates 1 or more correct conclusions; additional supporting computations beyond level 1 must be present. The work may contain some computational errors.

b) Student has correct mathematics for 1 or more options but indicates no conclusion.

3 a) Student shows correct mathematical evidence that Treena has 3 options, but the explanation is unclear or incomplete.

b) Student shows correct mathematical evidence for any 2 of Treena's 3 options and the explanation is clear and complete.

4 Full credit response - correct solution and complete, clear explanation

9 The work is completely incorrect, irrelevant, or off task. (Just computing 7 x 45 = 315 is a score of 9.)

0 = No response (blank)

Once the practice session was completed, the formal scoring process began. During the scoring, notes on various items were compiled for the readers for their reference and guidance. In addition, short training sessions were conducted when the team leaders determined by reviewing discrepancies that certain items were causing difficulties for the scorers.

During the first week of scoring, the team leaders reviewed 25 percent of the responses scored by each reader and brought any problems related to scoring to the attention of the individual reader. After the first week, leaders continued to review 25 percent of the booklets for any readers found to be having difficulties and 10 percent of the booklets scored by the rest of the readers. In this way, team leaders could be certain that their teams were scoring consistently. When a reader's score was judged to be discrepant with the scoring guides, the team leader discussed the response and its score with that reader. The team leaders also met as a group on a daily basis to discuss any problem responses to test items that had arisen in order to ensure that all teams of readers were scoring all items in exactly the same manner.

5.7.3 Trend Scoring of 1990 Items

During the scoring of the 1992 Trial State Assessment in mathematics, a trend scoring was conducted using a subsample of student test booklets from the mathematics portion of the 1990 NAEP. Four blocks of items used in the 1990 assessment were re-administered in 1992. Three of these blocks contained open-ended items (a total of 25 short constructed-response items). The training for the scoring of these items was conducted using training materials and scoring guides identical to those used for the 1990 assessment.

One hundred booklets from each of the 40 states that participated in the 1990 Trial State Assessment were chosen at random to be scored again in 1992. Each of these 4,000 booklets contained all three trend blocks. Scoring reliability for the 1990 trend scoring was 95.8 percent. This reliability percentage was calculated based on the rate of exact agreement between the scores given in 1990 and the scores given to the same student responses in 1992.

5.7.4 Reliability of Scoring

Twenty percent of the booklets containing constructed responses (for a total of over 6,500 responses) were scored by a second reader to obtain statistics on interreader reliability, which was determined by calculating the percent of exact agreement between readers. The overall interreader reliability for all constructed-response items combined was 94.1 percent. Reliabilities for the 11 extended constructed-response items ranged from 69.3 percent to 90.5 percent, with an average reliability of 81.1 percent (reliabilities for each of the 11 items are given in Table 5-1). For the 82 short constructed-response items, reliabilities ranged from 81.9 to 99.2 percent, with an average reliability of 96.2 percent. This reliability information was also used by the team leaders in monitoring the capabilities of all readers and the uniformity of scoring across readers.

Because the reliability scoring was done on separate scoring sheets, all reliability scoring was "blind," or uninfluenced by any score already given. The reliability scoring for each batch of

Table 5-1

Interreader Reliabilities for Extended Constructed-response Items in the 1992 Trial State Assessment in Mathematics

Grade 4

NAEP ID	Description	Content Area	Interreader Reliability
M041201	Compare two geometric shapes	Geometry	72.6
M043501	Explain solution to a problem involving counting	Algebra and Functions	89.5
M044401	Demonstrate understanding of place value ("Laura Use Calculator")	Numbers & Operations	90.5
M045401	Reason (meaning of fraction) ("Pizza Comparison")	Numbers & Operations	82.1
M049001	Identify correct pictograph ("Graphs of Pockets")	Data Analysis, Statistics, & Probability	76.8

Grade 8

NAEP ID	Description	Content Area	Interreader Reliability
M045901	Solve a problem involving intersecting circles ("Radio Stations")	Geometry	79.3
M051101	Reason to maximize difference	Numbers & Operations	69.3
M052201	Show how three figures can be divided to find area	Measurement	84.3
M053101	Find probability and explain	Data Analysis, Statistics, & Probability	85.1
M054301	Extend pattern to find term ("Marcy Dot Pattern")	Algebra & Functions	81.0
M055501	Plan and analyze expenses with given parameters ("Treena's Budget")	Numbers & Operations	81.5

booklets was completed by a team other than the one that did the full scoring. In this way, the full scoring and the reliability scoring was spread across all teams so that all readers were compared against all other readers.

5.8 DATA TRANSCRIPTION SYSTEMS

The transcription of the student response data into machine-readable form was achieved through the use of three separate systems: data entry (scanning), validation (pre-edit), and resolution.

5.8.1 Data Entry

The data entry process is the first time that booklet level data were input to the computer system. As all documents used in the 1992 assessment were scannable documents, the data were collected using NCS optical scanning equipment. The data were then edited and questionable data were resolved before further processing.

To ensure data integrity, edit rules were applied to each scanned data field. This procedure validated each field and reported all problems for subsequent resolution. After each field was examined and corrected, the edit rules were re-applied for final verification.

5.8.2 Scanning

After the initial manual verification, the scannable documents were transported to a slitting area where the folded and stapled spine was removed from each document. Scanning operations were performed by NCS's HPS Optical Scanning equipment. The optical scanning devices and software used at NCS permits a complete mix of NAEP scannable materials to be scanned with no special grouping requirements. However, for manageability and tracking purposes, student documents, excluded student questionnaires and teacher questionnaires were batched separately. In addition to the capture of scannable responses, the bar code identification numbers used to maintain process control were also decoded and transcribed to the NAEP computerized data file.

The scanning program is a table-driven software process that uses standard routines and application-specific tables to identify and define the documents and formats to be processed. When a booklet cover is scanned, the program uses the booklet number to determine the sequence of pages and the formats to be processed. By reading the booklet cover, the program recognizes which pages should follow and in what order.

The scanning program wrote four types of data records into the data set: a batch header record containing information coded onto the batch header sheet by receipt processing staff; a session header record containing information coded onto the session batch header sheet by receipt processing staff; a data record containing all of the translated marked ovals from all pages in a booklet; and a dummy data record, serving as a place holder in the file for a booklet with an unreadable cover sheet. The document code was written in the same location on all records to distinguish them by type.

The following coding rules were used:

- The data values from the booklet covers and scorer identification fields were coded as numeric data.
- Unmarked fields were coded as blanks and processing staff were alerted to missing or uncoded critical data.
- Fields that had multiple marks were coded as asterisks (*).
- The data values for the item responses and scores were returned as numeric codes.
- The multiple-choice, single-response format items were assigned codes depending on the position of the response alternative; that is, the first choice was assigned a 1, the second a 2, and so forth.
- The circle-all-that-apply items were given as many data fields as response alternatives; the marked choices are coded as 1 and the unmarked choices as blanks.
- The fields from unreadable pages were coded with an X as a flag for resolution staff to correct.

5.9 DATA VALIDATION

The data entry and resolution system used for the Trial State Assessment Program was also used for the national assessment program. The system is able to process materials submitted from both scannable and nonscannable media simultaneously for three age groups, three assessment types, and five questionnaires. The use of batch identification codes—comprising the school and session codes as well as the batch sequence numbers for suspect record identification—facilitated the management of the system and correction of incorrectly gridded or keyed information.

As the program processed each data record, it first read the booklet number and checked it against the batch session code for appropriate session type. Any mismatch was recorded on the error log and processing continued. The booklet number was compared against the first two digits of the student identification number. If they disagreed, because of improper bar coding, a message was written to the error log. The remaining booklet cover fields were then read and validated for the correct range of values. The school codes had to be identical to those on the Process Control System record and the grade code had to be either 4 or 8. All data values that were out of range were read as is, but flagged as suspect. All data fields that were read as asterisks were recorded on the edit log.

Document definition files describe each document as a series of blocks that are described as a series of items. The blocks in a document were traversed in the order that they appear on the document. Each block's fields were validated during this process. If a document contained suspect fields, the cover information was recorded on the edit log with a description of the suspect data. Some fields (e.g., AGE or DOB), required special types of edits. These fields were identified in the document definition fields, and a subroutine was invoked to handle these cases.

The program next cycled through the data area corresponding to the item blocks. The task of translating, validating, and reporting errors for each data field in each block was performed by a routine that required only the block identification code and the string of input data. This routine had access to a block definition file that had the number of fields to be processed for each block and the field type (alphabetic or numeric), the field width in the data record, and the valid range of values for each field. The routine processed each field in sequential order, performing the necessary translation, validation, and reporting tasks.

The first of these tasks checked for the presence of blanks or asterisks in a critical field. These were recorded on the edit log and processing continued with the next field. No action was taken on blank-filled fields for multiple-choice items since that code indicated a nonresponse. The field was validated for range of response, recording anything outside of that range to the edit log. The item type code was used by the program to make a further distinction among constructed-response item scores and other numeric data fields. Moving the translated and edited data field into the output buffer was the last task performed in this phase of processing.

The completed string of data was written to the data file when the entire document had been processed. Then, when the next session header record was encountered, the program repeated the same set of processes for that session. The program closed the data set and generated an edit listing when it encountered the end of a file.

Accuracy checks were performed on each batch processed. Every 500th document of each booklet form was printed in its entirety, with a minimum of one document type per batch. This record was checked, item by item, with the source document for errors.

5.10 EDITING

Quality procedures and software throughout the system ensure that the NAEP data are correct. The initial editing that took place during the receipt control process included verification of the schools and sessions. Receipt control personnel checked that all student documents on the Administration Schedule were undamaged and assembled correctly. The machine edits performed during data capture verified that each sheet of each document was present and that each field had an appropriate value. All batches entered into the system were edited for errors.

Data editing occurred after these checks and consisted of a computerized edit review of each respondent's document and the clerical edits necessary to make corrections based upon the computer edit. This data editing step was repeated until all data were correct.

The first phase of data editing was designed to ensure that all documents were present. A computerized edit list was produced after NAEP documents were scanned and with the supporting documentation sent from the field the edit function was performed. The hard copy edit list contained all the vital statistics about the batch and each school and session within the batch, such as the number of students, school code, type of document, assessment code, error rates, suspect cases, and record serial numbers. Using these inputs, the data editor verified that the batch had been assembled correctly, each school number was correct, and all student documents within each session were present.

During data entry, counts of documents processed by type were generated. These counts were checked against the Administration Schedule counts entered into the Process Control System during the receiving process. The number of assessed and absent students processed had to match the number of used booklets indicated on the Process Control System.

The second phase of data editing was carried out by an experienced editing staff using a predetermined set of rules to review the field errors and record corrections to be made to the student data file. The same computerized edit list used in the first phase was also used to perform this function.

The editing staff made corrections using the edit log prepared by the computer and the actual source document listed on the edit log. The corrections were identified by batch sequence numbers and field name for suspect record and field identification. The edit log indicated the current composition of the field. This particular piece of information was then visually checked against the NAEP source document by the editing staff for double grids, erasures, smudge marks or omitted items that were flagged. Each flagged item was handled in one of the following ways:

- *Correctable Error:* If the error could be corrected by the editing staff, according to the editing specifications, the corrections were indicated on the edit listing.

- *Field Correctable:* If an error was not correctable according to the specifications, an alert was issued to the operations coordinator for resolution. Once the correct information was obtained, the correction was indicated on the edit listing.

- *Noncorrectable Error:* If an error suspect was found to be correct as stated, and no alteration was possible according to source documents and specifications, the programs were tailored to allow this information to be accepted into the data record and no corrective action was taken.

These corrections were noted on the edit list. When the entire batch of sessions was resolved, the list was forwarded to the key entry staff. The corrections were entered and verified through the Falcon system. When all corrections were entered and verified for a batch, an extract program was run to pull the correction records to a mainframe data set.

The post-edit program was initiated next. This program applied the corrections to the specified records and once again applied the error criteria to all records. If there were further errors, another edit list was printed and the cycle began again.

When the edit process had produced an error-free file, the booklet ID number was posted to the NAEP tracking file by school and sessions. This allowed for an accumulation process to accurately measure the number of documents processed for a session within a school and the number of documents processed by form. The posting of booklet IDs also ensured that a booklet ID was not processed more than once. These data allowed the progress of the assessment to be monitored and reported on the status report.

At this point, a job was automatically submitted to produce the NAEP scoring sheets for this batch. The program also selected the records to be scored by a second reader for reliability. These sheets were printed, matched with the original documents, and forwarded to the NAEP scoring area.

Once all documents for a batch had been scored, the sheets were batched and submitted to scanning. A series of edits were run to verify the information on these sheets. The scorer identification fields were processed at this point and certain checks were made. The routine validated the score range and did not permit a blank field. If no score was indicated or the score was out of range, the disparity was noted on the edit log.

These error logs were returned to the scoring groups for resolution and the corrections were entered directly to the files. The edit process was repeated until the file was error free.

As a final quality control check, ETS identified a random sample of each booklet type from the master student file. The designated documents and scoring sheets were located, removed from storage and forwarded to ETS for quality control (see Chapter 6). On completion of quality control processing, the booklets were returned to NCS for return to storage.

5.11 QUESTIONNAIRES

The questionnaires were received either with the session shipment or in a later shipment. Once the questionnaires were verified with the roster, they were accumulated by the receiving clerks. The school characteristics and policies questionnaires, teacher questionnaires and excluded student questionnaires were batched and sent to scanning at regular intervals. Every effort was made to keep current on all forms, both to ensure the processing of all documents for a session and to deliver all data at the same time.

All documents, regardless of method of entry, were run through the process of error identification and resolution.

5.12 MERGING OF STUDENT DATA

At the completion of the scoring and verification of the constructed responses, the complete records for students were merged. This merge included the machine-scanned data, the scores to the constructed responses, and the responses from the estimation booklets. Verification of complete student records was conducted prior to the delivery of the data files.

5.13 STORAGE OF DOCUMENTS

Once the editing process had been successfully completed on the batches, they were sent to the NCS warehouse for storage. The storage location of all documents was recorded on the inventory control system and stored for later retrieval. Unused materials were sent to temporary storage until the completion of the assessment and acceptance of the data files, at which time they were destroyed.

Chapter 6

CREATION OF THE DATABASE AND EVALUATION OF THE QUALITY CONTROL OF DATA ENTRY

John J. Ferris and David S. Freund

Educational Testing Service

6.1 OVERVIEW

The data transcription and editing procedures described in Chapter 5 resulted in the generation of disk and tape files containing various data for assessed students, excluded students, teachers, and schools. The weighting procedures described in Chapter 7 resulted in the generation of data files that included the sampling weights required to make valid statistical inferences about the population from which the 1992 fourth- and eighth-grade Trial State Mathematics Assessment samples were drawn. These files were merged into a comprehensive, integrated database. To evaluate the effectiveness of the quality control of the data entry process, the final integrated database was sampled, and the data were verified in detail against the original instruments received from the field.

This chapter begins with a description of the transcribed data files and the procedure of merging them, or bringing them together, to create the 1992 Trial State Assessment database for fourth- and eighth-grade students. The last section presents the results of the quality control evaluation.

6.2 MERGING FILES INTO THE TRIAL STATE ASSESSMENT DATABASE

The transcription process conducted by National Computer Systems resulted in the transmittal to ETS of four data files for both fourth and eighth grade: one file for each of the three questionnaires (teacher, school, and excluded student) and one for the student response data. The sampling weights, derived by Westat, Inc., comprised an additional three files for each grade—one for students, one for schools, and one for excluded students. (See Chapter 7 for a discussion of the sampling weights.) These seven files at each grade were the foundation for the analysis of the 1992 Trial State Assessment data. Before data analyses could be performed, these data files had to be integrated into a coherent and comprehensive database.

The 1992 Trial State Assessment database for fourth and eighth grade consisted of three files—student, school, and excluded student. Each record on the student file contained a student's responses to the particular assessment booklet the student was administered (booklets 1 to 26), the student's responses to booklet 29 (a single block of estimation items that was administered to all assessed students), and the information from the questionnaire that the student's mathematics teacher completed. (See Chapter 2 for information regarding assessment instruments.) Since teacher response data can be reported only at the student level, it was not necessary to have separate teacher files. The school files and excluded student files were separate and could be linked to the student files through the state and school codes.

The creation of the student data files for fourth and eighth grade began with the reorganization of the data files received from National Computer Systems. This involved two major tasks: 1) the files were restructured, eliminating unused (blank) areas to reduce the size of the files; and 2) in cases where students had chosen not to respond to an item, the missing responses were recoded as either "omit" or "not reached," as appropriate. Next, the student response data were merged with the student weights file. The resulting file was then merged with the teacher response data. In both merging steps, the booklet ID (the two-digit booklet number and a five-digit serial number) was used as the matching criterion.

The school file for each grade was created by merging the school questionnaire file with the school weights file and a file of school variables, supplied by Westat, which included demographic information about the schools collected from the principal's questionnaire. The state and school codes were used as the matching criteria. Since some schools did not return a questionnaires and/or were missing principal's questionnaire data, some of the records in the school file contained only school-identifying information and sampling weight information.

The excluded student file for each grade was created by merging the excluded student questionnaire file with the excluded student weights file. The assessment booklet serial number was used as the matching criterion.

When the student, school, and excluded student files for each grade had been created, the database was ready for analysis. In addition, whenever new data values, such as composite background variables or plausible values, were derived, they were added to the appropriate database files using the same matching procedures as described above.

For archiving purposes, restricted-use data files and codebooks for each state were generated from this database. The restricted-use data files contain all responses and response-related data from the assessment, including responses from the student booklets and teacher and school questionnaires, proficiency scores, sampling weights, and variables used to compute standard errors.

6.3 CREATING THE MASTER CATALOG

A critical part of any database is its processing control and descriptive information. Having a central repository of this information, which may be accessed by all analysis and reporting programs, will provide correct parameters for processing the data fields and consistent labeling for identifying the results of the analyses. The Trial State Assessment master catalog file was designed and constructed to serve these purposes for the Trial State Assessment database.

Each record of the master catalog contains the processing, labeling, classification, and location information for a data field in the Trial State Assessment database. The control parameters are used by the access routines in the analysis programs to define the manner in which the data values are to be transformed and processed.

Each data field has a 50-character label in the master catalog describing the contents of the field and, where applicable, the source of the field. The data fields with discrete or categorical values (e.g., multiple-choice items and professionally scored items, but not weight fields) have additional label fields in the catalog containing 8- and 20-character labels for those values.

The classification area of the master catalog record contains distinct fields corresponding to predefined classification categories (e.g., mathematics content and process areas) for the data fields. For a particular classification field, a nonblank value indicates the code of the subcategory within the classification categories for the data field. This classification area permits the grouping of identically classified items or data fields by performing a selection process on one or more classification fields in the master catalog.

The master catalog file was constructed concurrently with the collection and transcription of the Trial State Assessment data so that it would be ready for use by analysis programs when the database was created. As new data fields were derived and added to the database, their corresponding descriptive and control information were entered into the master catalog.

6.4 QUALITY CONTROL EVALUATION

The purpose of the data entry quality control procedure is to gauge the overall accuracy of the process that transforms responses into machine-readable data. The procedure involves examining the actual responses made in a random sample of booklets and comparing them with the responses recorded in the final database, which is used for analysis and reporting.

6.4.1 Student Data

Twenty-six assessment booklets numbered 1 through 26 and an estimation block identified as booklet 29 were administered as part of the Trial State Assessment in mathematics. Table 6-1 provides the numbers of each booklet at each grade for which data were scanned into data files. These numbers varied somewhat more than in the 1990 assessment, but chi-square measures of the variation proved to be insignificant at both grades.

Since booklet 29 was administered to all students, it was treated for quality control purposes as an extension of each of booklets 1 through 26. All data for a selected student were collected and examined, including data from booklet 29.

The number of students assessed in each of the 44 participating jurisdictions varied also. At grade 4, 29 jurisdictions met or exceeded the target of 2,500 students and a few smaller jurisdictions fell several hundred short of the target. The average number of fourth-grade students assessed in each jurisdiction was 2,523. At grade 8, 26 jurisdictions met or exceeded

Table 6-1

Number of Assessment Booklets Scanned and Selected for Quality Control Evaluation

Booklet Number	Total Booklets Scanned		Total Booklets Selected	
	Grade 4	Grade 8	Grade 4	Grade 8
1	4,211	4,129	11	13
2	4,186	4,142	12	10
3	4,162	4,140	8	9
4	4,169	4,167	13	7
5	4,151	4,165	11	8
6	4,227	4,206	10	9
7	4,211	4,185	11	13
8	4,192	4,198	12	13
9	4,279	4,177	12	11
10	4,323	4,234	10	9
11	4,377	4,258	9	13
12	4,332	4,221	10	10
13	4,388	4,203	11	7
14	4,363	4,163	10	12
15	4,338	4,151	10	13
16	4,375	4,155	12	15
17	4,371	4,179	11	8
18	4,331	4,186	13	13
19	4,334	4,210	8	8
20	4,329	4,192	11	10
21	4,289	4,225	11	11
22	4,234	4,174	10	10
23	4,251	4,156	10	9
24	4,221	4,125	4	9
25	4,187	4,135	11	13
26	4,161	4,110	10	8
Total	110,992	108,586	271	271

target of 2,500 students, the same jurisdictions as at grade 4 fell short, and the average number of students assessed was 2,468. To simplify the selection of booklets for examination, a method was developed that involved selecting all occurrences of a specified booklet in a randomly selected "stack." A stack is a unit of collection containing anywhere from 11 to 105 booklets, but typically between 50 and 60 booklets, in an assortment related to the spiraling technique used to distribute the booklets. The selection method was designed to yield approximately the same number of each booklet but, due to the variability in the size and contents of the stacks, there was somewhat more variation in the numbers of booklets selected than in the 1990 assessment (see Table 6-1). However, all of the booklets were sampled in adequate numbers and the average rate of selection was 1/410 at grade 4 and 1/401 at grade 8, a selection rate comparable to that used in past assessments at both the state and national levels. The few errors found during this quality control examination did not cluster by booklet number, so there is no reason to believe that the variation in numbers of booklets selected had a significant effect on the estimates of overall error rate confidence limits reported below.

The quality control evaluation detected only 10 errors total for both grades in these booklet samples—six instances of multiple responses that were not identified as such by the scanner, and four instances of erasures that were recorded instead of ignored. The usual quality control analysis based on the binomial theorem permits the inferences described in Table 6-2.

Table 6-2
Inference from the Quality Control Evaluation of Student Data

Subsample	Entry Type	Different Booklets Sampled	Number of Booklets Sampled	Characters Sampled	Number of Errors	Observed Rate	99.8% Confidence Limit
Grade 4	Scanned	26	271	33,682	1	.00003	.0003
Grade 8	Scanned	26	271	38,888	9	.00020	.0006

The grade 8 error rate is about the same as was observed in the 1990 assessment. For some reason, the students in grade 4 did not seem to challenge the scanner with erasures and optical ambiguities, so the error rate confidence limit for that grade was much lower. Neither error rate offers the threat of interference with the validity of any data analyses. As usual, there was some indication that the error rates could be improved with further tuning of the scanner procedures, but the process as it stands can certainly be described as clean and reliable. A very large volume of data was scanned with consistently excellent results.

6.4.2 Teacher Questionnaires

A total of 14,553 questionnaires at grade 4 and 11,453 questionnaires at grade 8 were collected from mathematics teachers. Questionnaires were sampled at the rate of 1 in 200, resulting in the selection of 72 questionnaires at grade 4 and 58 questionnaires at grade 8. The selected questionnaires contained a total of 13 errors, usually involving the scanner's mistaking an erasure for a response, but occasionally involving the failure of the scanner to pick up a multiple response. In every case, the respondent's intention was clear to the human eye, but the scanner seemed unprepared to exercise the same judgment that a careful observer would. The result is an error rate for the teacher questionnaire data that is 5 to 10 times as high as for the student data. One possible explanation for this is that teacher questionnaires are inherently more complex than student assessment booklets, which leads to a much higher rate of erasures and other errors by the respondents. Perhaps a redesign of these questionnaires would bring the error rate down. This is not to say that the degree of erroneous data in the teacher questionnaire files is worrisome, but rather that the student data are much more error-free. There is every indication that the quality of the teacher data is more than adequate for the purposes to which it was put.

6.4.3 School Questionnaires

A total of 4,857 questionnaires at grade 4 and 3,699 at grade 8 were collected from school administrators. These questionnaires were sampled for quality control evaluation at the rate of 1 in 50, resulting in the selection of 97 questionnaires at grade 4 and 74 at grade 8. In the 1990 assessment, data from the school questionnaires had been key-entered; for the 1992 assessment, the completed questionnaires were machine-scanned. It is interesting to compare these two very different data entry methods. While the overall error rates were the same, none of the errors in the keyed data involved the misreading of erasures or multiple responses; in the scanned data, all of the errors were of this type.

Again, the quality of the data was very good, with an error rate of about half that of the teacher questionnaire data.

6.4.4 Excluded Student Questionnaires

A total of 13,268 excluded student questionnaires were scanned at grade 4, and 6,454 at grade 8. These were sampled at the rate of about 1 in 200, resulting in the selection of 66 questionnaires at grade 4 and 33 questionnaires at grade 8. All the errors found were due to the scanner's mistaking an erasure for an intended response.

The quality of these data appears to be about as high as the other questionnaires—that is to say, adequate for the purposes to which it was put. The results of the evaluation of the questionnaire data are summarized in Table 6-3.

Table 6-3

Inference from the Quality Control Evaluation of Questionnaire Data

Subsample*	Entry Type	Different Booklets Sampled	Number of Booklets Sampled	Characters Sampled	Number of Errors	Observed Rate	99.8% Confidence Limit
Grade 4							
TQ	Scanned	1	72	8,136	9	.0011	.0026
SQ	Scanned	1	97	9,312	4	.0004	.0015
XQ	Scanned	1	66	5,148	4	.0008	.0027
Grade 8							
TQ	Scanned	1	58	6,264	4	.0006	.0022
SQ	Scanned	1	74	7,104	2	.0003	.0012
XQ	Scanned	1	33	2,574	3	.0012	.0047

---

* TQ = Teacher questionnaire; SQ = School questionnaire; XQ = Excluded student questionnaire

Chapter 7

WEIGHTING PROCEDURES AND VARIANCE ESTIMATION

Adam Chu and Keith F. Rust

Westat, Inc.

7.1 INTRODUCTION

Following the collection of assessment and background data from and about assessed and excluded students, sampling weights and associated sets of replicate weights were derived. The sampling weights are needed to make valid inferences from the student samples to the respective populations from which they were drawn. Replicate weights are used in the estimation of sampling variance, through the procedure known as jackknife repeated replication.

Each student was assigned a weight to be used for making inferences about the state's students. This weight is known as the *full-sample* or *overall* sample weight. In the 1990 Trial State Assessment Program, a second weight, known as the comparison weight, was also derived for the purpose of comparing the assessment performance of students in monitored sessions with those in unmonitored sessions. However, for the 1992 Trial State Assessment Program, comparison weights were not calculated. Valid (i.e., unbiased) comparisons of this kind can be made using the full sample weights; however, the standard errors associated with these comparisons are somewhat larger than those that would be obtained using comparison weights.

The full-sample weight contains three components. First a base weight is established that is the inverse of the overall probability of selection of the sampled student. The base weight incorporates the probability of selecting a school and the student within a school, and accounts for the impact of procedures used to keep to a minimum the overlap of the state school sample with the NAEP national sample and the sample for the National Longitudinal Study of Chapter 1 Children (see Chapter 3). The base weight is then adjusted for two sources of nonparticipation—school-level and student-level. These weighting adjustments seek to reduce the potential for bias from such nonparticipation by increasing the weights of students from schools similar to those schools not participating, and increasing the weights of students similar to those students from within participating schools who did not attend the assessment session (or a makeup session) as scheduled. The details of how these weighting steps were implemented are given in sections 7.2 and 7.3.

In addition to the full-sample estimation weights, a set of replicate weights was provided for each student. These replicate weights are used in calculating the sampling errors of estimates obtained from the data, using the jackknife repeated replication method. Full details of the method of using these replicate weights to estimate sampling errors are contained in the technical reports for the 1988 and 1990 national assessments (Johnson & Zwick, 1990; Johnson & Allen, 1992). Section 7.5 of this report describes how the sets of replicate weights were generated for the 1992 Trial State Assessment data. The methods of deriving these weights were aimed at reflecting the features of the sample design appropriately in each state, so that when the jackknife variance estimation procedure is implemented, approximately unbiased estimates of sampling variance result.

7.2 CALCULATION OF BASE WEIGHTS

The base weight assigned to a school was the reciprocal of the probability of selection of that school. For the fourth-grade samples, the school base weight depended on the subject of assessment since some schools were so small that students were tested in only one subject in those schools. Under the sample selection procedures used for the 1992 Trial State Assessment Program (see Chapter 3), the school selection probability may be greater than 1 for large schools. In this case, the probability of selection actually represents the expected number of times the school would be selected under the systematic sampling process. In general, the school base weight reflected the actual probability used to select the school from the frame, including the impact of avoiding schools selected for the NAEP national sample and the sample for the National Longitudinal Study of Chapter 1 Children (see Chapter 3).

The student base weight was obtained by multiplying the school base weight by the within-school student weight, where the within-school student weight reflected the probability of selecting students within the school for a particular assessment subject. Additional details about the weighting process are given in the sections below.

7.2.1 Calculation of School/hit Base Weights

As described in section 3.4.5, schools were sometimes selected in clusters in order to avoid giving small schools an extremely low probability of selection. Moreover, large clusters (or schools) could have been selected more than once in the systematic sampling process. If a large cluster (or school) was selected more than once, each selection or "hit" was treated separately in the selection of students within a school. For example, a school that was selected twice was allocated twice the usual numbers of students for the assessments; a school that was selected three times was allocated three times the usual numbers of students for the assessments.

The weight for sample cluster c was computed as:

$$W_c^{clus} = \frac{E}{mE_c}$$

where

E_c = the enrollment in the given grade for the *c*th cluster in the state;

$$E \quad = \quad \sum_{c=1}^{M} E_c$$

$$= \quad \text{the state-wide enrollment in the given grade; and}$$

m = the number of cluster/hits selected from the state.

If a cluster was selected more than once, each hit received the same base weight. In general, the base weight for sample school i (or school/hit i if the school was selected more than once) in a given state was computed as:

$$W_i^{sch} = W_{ci}^{clust} T_{ci}$$

where W_{ci}^{clust} is the base weight of the cluster containing school i and T_{ci} is a "thinning" factor that reflects the fact that small schools in the Cluster Type 2 states were subject to thinning (see section 3.5.3). The thinning factor T_{ci} was equal to the ratio of the sampling size measure of the largest school in the cluster to the size measure of the retained school.

Since all schools in Cluster Type 1 states were included in the sample with certainty (see section 3.5.2), they were assigned school base weights (W_i^{sch}) equal to 1.

7.2.2 Weighting New Schools

As described in Chapter 3, new schools were sampled from the updated sampling frame list from each district in a sample of districts. In a few states, the selection probabilities of some new schools were quite small, resulting in excessively large school base weights. Where the weighted contribution to the estimate of total enrollment of a new school exceeded three times the median contribution, the base weight for that school was adjusted downwards (trimmed) in order to reduce the impact of the extreme weights on the variance of the estimates. Base weights were trimmed for a total of nine new schools in the following six states: New Jersey (grades 4 and 8), North Carolina (grade 4), Indiana (grade 8), Kentucky (grade 8), New York (grade 8), and Ohio (grade 4). For these nine schools, the trimmed school weight (which was then used in the subsequent calculation of nonresponse adjustments) was computed as:

$$W_i^{sch} = \frac{E_{max}}{E_i}$$

where E_i is the estimated grade enrollment of the new school, and E_{max} is the maximum allowable weighted contribution to the estimated total grade enrollment for the given state. The value of E_{max} was established so that the weighted contribution of the new school to the total weighted grade enrollment never exceeded about three times the median value of the distribution of weighted enrollment counts for the remaining schools in the sample.

This adjustment was made to avoid introducing substantial variability into the sample estimates, as a result of giving relatively very large weights to one or two schools, and thus the sampled students within them. Although this procedure technically introduces a bias in the estimates for these states, we judged that it would be trivial in comparison to the level of sampling variance. For a discussion of issues involved in trimming of survey weights, see Potter (1988) and Stokes (1990).

7.2.3 Treatment of Substitute and Double-session Substitute Schools

Schools that replaced a refusing school (i.e., substitute schools) were assigned the weight of the refusing school, unless the substitute school also refused. Schools conducting extra sessions that served as substitutes for a refusing school (i.e., double-session substitutes) in effect had two school weights. The students in the school who were assigned to the original session were given the school base weight of the participating school, while those students assigned to the extra session(s) were assigned the school base weight of the refusing school.

7.2.4 Calculation of Student Base Weights

Within the sampled schools, eligible students were assigned to sessions using the procedures described in sections 3.5.7 and 3.6. The within-school probability of selection for assessment in mathematics therefore depended on the number of grade-eligible students in the school and the number of students selected for the assessment (usually 30). The within-school weights for the substitute schools were further adjusted to compensate for differences in the sizes of the substitute and the originally sampled (replaced) schools. In the case of the fourth-grade sample, the within-school weight also reflected the fact that a small school could have been selected for one subject but not the other. Thus, in general, the within-school student weight for the *j*th student in school *i* was equal to:

$$W_{ij}^{within} = \frac{N_i}{n_i} K_{1i} K_{2i}$$

where

N_i = the number of grade-eligible students enrolled in the school as reported in the sampling worksheets; and

n_i = the number of students selected for the given subject.

The factors K_{1i} and K_{2i} in the formula for the within-school student weight generally apply to only a few schools in each state. The factor K_{1i} adjusts the count of grade-eligible students in a substitute school to be consistent with corresponding count of the originally sampled (replaced) school. Specifically, for substitute schools,

K_{1i} = $\frac{E_i}{E_i^{su}}$

E_i = the QED grade enrollment of the originally sampled (replaced) school; and

E_i^{su} = the QED grade enrollment of the substitute school.

For nonsubstitute schools, $K_{1i} = 1$.

The factor K_{2i} applies only to the fourth-grade sample and reflects the subsampling procedure used to select the subject in which students in small schools were to be assessed (section 3.5.7). For a given subject, K_{2i} is defined as follows:

$$K_{2i} = \begin{cases} 1 & \text{if the fourth-grade school was selected for both subjects;} \\ 2 & \text{if the fourth-grade school was selected only for the given subject} \\ 0 & \text{if the fourth-grade school was not selected for the given subject} \end{cases}$$

Note that if K_{2i} is 2 for mathematics (say), then K_{2i} is 0 for reading, and vice versa.

The overall student base weight for a student j selected for mathematics assessment in school i was then computed as:

$$W_{ij}^{base} = W_i^{sch} W_{ij}^{within} \; .$$

Checks were made on these student base weights to ensure that the value was always 1.0 or greater.

7.3 Adjustments for Nonresponse

The base weight for a student was adjusted by two factors: one to adjust for nonparticipating schools for which no substitute participated, and one to adjust for students who were invited to the assessment but did not appear in either the scheduled or makeup sessions.

7.3.1 Defining Initial School-level Nonresponse Adjustment Classes

School-level nonresponse adjustment classes were initially created based on the urbanicity and minority strata used in sampling. In states and urbanicity strata where minority stratification was not used, nonresponse classes were created based on median household income.

The procedure for creating income classes was as follows. First, three classes of schools were formed for each urbanicity stratum so that (1) each class had approximately the same number of sample schools and (2) the classes were ranked from low to high income. This was done using only the schools in the sample (including new schools), sorting them by median income, and then dividing the schools into three groups with equal numbers of schools. In a few states (Cluster Type 3 states) only large schools (those with grade enrollment over 20) were used to form the income strata, although all schools were classified into either income or minority strata. In creating the nonresponse adjustment classes, urbanicity was used as the primary variable and minority/income was used as the secondary variable.

The initial nonresponse adjustment classes (i.e., sampling strata) are summarized for each state in Tables 3-3 and 3-4 of Chapter 3. As can be seen in these tables, the definition of the initial nonresponse adjustment classes varied from one state to another. For example, nine classes obtained by cross-classifying three levels of urbanicity (central city, suburban, other) with three levels of minority status (low, medium, and high) were defined for Alabama, whereas for New York, the classes were defined by minority status within the central city and suburban strata, and by income classes within the rural stratum.

7.3.2 Constructing the Final Nonresponse Adjustment Classes

The objective in forming the final nonresponse adjustment classes was to create as many classes as possible that were internally as homogeneous as possible, but such that the resulting nonresponse adjustment factors were not subject to large random variation. The procedures discussed below were established with the aim of meeting this objective.

The schools (or school/hits in the case of schools that were selected more than once in the sampling process) were sorted into the initial nonresponse classes defined in Tables 3-3 and 3-4 and the following unweighted and weighted counts and ratios were produced for each class:

- total in-scope school/hits from the original sample (an in-scope school is one that has at least one eligible student enrolled);
- participating in-scope schools from the sample (both original and substitutes); and
- total in-scope schools from the original sample divided by participating in-scope schools from the sample

The weights used in the calculations were the school/hit base weights defined in section 7.2, multiplied by the QED grade enrollment for the school.

The following guidelines were established for reviewing these counts and ratios and determining what collapsing should be done. Within an initial nonresponse class, if the weighted ratio of in-scope schools to participating schools was less than 1.35, with at least six participating schools in the class, there was no need to collapse the particular cell. If any nonresponse class had fewer than 6 schools or a ratio greater than or equal to 1.35, it was collapsed with another class such that the new class met these conditions. The order of variables to be collapsed (from most desirable to least desirable) was income strata or minority strata, followed by urbanicity strata. The exceptions occurred in cases where minority classes within an urbanicity stratum varied considerably as to the relative sizes of the minority population. In such cases, we collapsed over urbanicity first to keep the classes as homogeneous as possible with regard to race/ethnicity. In some cases, final classes were formed with ratios in excess of 1.35. This occurred in states with relatively high school nonresponse. In no case was a class formed with fewer than six school/hits.

The choices of 1.35 as a cutoff for the nonresponse adjustment and 6 as the minimum number of participants within a class were both motivated by the desire to balance two conflicting needs. These are described in the first paragraph of this section. These limits were chosen on the basis of practical experience, combined with the application of theory about the effects of nonresponse class size on the accuracy of survey estimates, in a manner appropriate for the levels of nonresponse encountered in the various states.

7.3.3 School/hit Adjustment Factors

The school-level nonresponse adjustment factor for the *i*th school/hit in the *h*th class was computed as:

$$F_h^{(1)} = \frac{\sum_{i \in C_h} W_{hi}^{sch} E_{hi}}{\sum_{i \in C_h} W_{hi}^{sch} E_{hi} \delta_{hi}}$$

where

C_h = the subset of school/hit records in class *h*;

W_{hi}^{sch} = the base weight of the *i*th school/hit in class *h*;

E_{hi} = the QED grade enrollment for the *i*th school/hit in class *h*;

$$\delta_{hi} = \begin{cases} 1 & \text{if the } i\text{th school/hit in adjustment class } h \text{ participated in the assessments; and} \\ 0 & \text{otherwise.} \end{cases}$$

In the calculation of the above nonresponse adjustment factors, a school was said to have participated if

- it was selected for the sample from the QED frame or from the lists of new schools provided by participating school districts, and student assessment data were obtained from the school;

- the school refused but was replaced by a regular substitute school and student assessment data were obtained from the substitute school (so that the substitute participated in place of the originally selected school); or

- the school refused but was replaced by a double-session substitute school and the double-session substitute provided student assessment data for both the original and substitute sessions (so that the substitute school conducted additional sessions to replace the originally selected school).

Both the numerator and denominator of the nonresponse adjustment factor contained only in-scope schools.

The nonresponse-adjusted weight for the *i*th school/hit in class *h* was computed as:

$$W_{hi}^{adj} = F_h^{(1)} W_{hi}^{sch} \quad .$$

7.3.4 Student-level Nonresponse Adjustment Classes

The variables used to define initial classes for adjusting for student nonresponse were:

- the final school-level nonresponse adjustment classes described in section 7.3.2;

- the age class of the student; and

- the monitor status of the session the student attended.

Two age classes, "old" and "young," were defined for both grades. For grade 8, the "old" students were those born in September 1977 or earlier, while the "young" students were those born after September 1977. For grade 4, "old" students were those born in September 1981 or earlier; "young" students were those born after September 1981. Students in the "old" class are to some extent outliers with regard to age among their cohort. Previous findings from NAEP have shown that students in the "old" group tend to have higher absentee rates and lower proficiency scores than do students in the "young" group.

In order to determine whether the initial nonresponse classes needed collapsing, we reviewed the unweighted and weighted counts of assessed and absent students in each initial cell. (Excluded students were processed separately, using essentially the same procedures developed for assessed students.) The weight used for each student was the student base weight, adjusted for school nonresponse ($W_{hij}^{(2)}$ in section 7.3.5). The following guidelines were established for collapsing the initial nonresponse cells when necessary. Any cell with fewer than

20 assessed students was collapsed regardless of the value of the adjustment factor. If a cell had between 20 and 30 assessed students and the ratio of the weighted count of invited students to the weighted count of assessed students was greater than 1.5, the cell was collapsed. If a cell had more than 30 assessed students and the ratio of the weighted count of invited students to the weighted count of assessed students was greater than 2.0, the cell was collapsed.

When necessary, the collapsing of the initial cells proceeded as follows: First, collapsing was done across monitor status within all other classes. If the resulting cell still needed to be collapsed, the collapsing across monitor status was undone, and new cells were formed by collapsing across minority/income class. If these new cells still needed to be collapsed, collapsing across monitor status was done, followed by collapsing by urbanicity class and finally by age group, if necessary. Based on these guidelines, some collapsing was done for all states, usually over monitor status and particularly for "old" students.

7.3.5 Student Nonresponse Adjustments

As described above, the student-level nonresponse adjustments for the assessed students were made within classes defined by the final school-level nonresponse adjustment cells, monitor status of the school, and age group of the students. Let the *k*th final (collapsed) nonresponse class be denoted as A_k. The adjusted student base weight for the *j*th sample student in school/hit *i* in class A_k was calculated as:

$$W_{hij}^{(2)} = W_{hi}^{adj} W_{ij}^{within} = W_{hij}^{base} F_k^{(1)}$$

where

W_{hi}^{adj} = the nonresponse-adjusted school/hit weight for school/hit *i* in school adjustment class *h*;

W_{ij}^{within} = the within-school weight for the *j*th student in school *i*;

W_{hij}^{base} = $W_{hi}^{adj} W_{ij}^{within}$

= the student base weight for student *j* in school *hi*.

Using the adjusted student base weights, the assessed student nonresponse adjustment was calculated within nonresponse adjustment class A_k as:

$$F_k^{(2)} = \frac{\sum_{j \in A_k} W_{hij}^{(2)}}{\sum_{j \in A_k} W_{hij}^{(2)} \delta_{ij}}$$

where

$$\delta_{kj} = \begin{cases} 1 & \text{if the } j\text{th student in adjustment class } k \text{ participated in the assessments; and} \\ 0 & \text{otherwise.} \end{cases}$$

For excluded students, the same basic procedures as described above for assessed students were used, except that the numerator and denominator contained excluded rather than assessed students, and monitor status and student age group were not used to form the adjustment classes.

The final student weight for the jth student in class k was then computed as:

$$W_{kj}^{final} = F_k^{(2)} W_{kj}^{(2)}$$

Tables 7-1 and 7-2 summarize the final unweighted and weighted counts of assessed and excluded students, by state and grade. Checks were made on the final student weight distributions and totals at the state and subgroup within state, to ensure that there were no unexpected weight outliers or unusual distributions.

7.4 Characteristics of Nonresponding Schools and Students

In the previous section procedures were described for adjusting the survey weights so as to reduce the potential bias of nonparticipation of sampled schools and students. To the extent that a nonresponding school or student is different from those respondents in the same nonresponse adjustment class, potential for nonresponse bias remains.

In this section, we examine the potential for remaining nonresponse bias in two, related, ways. First we examine the weighted distributions, within each grade and state, of certain characteristics of schools and students, both for the full sample and for respondents only. This analysis is of necessity limited to those characteristics that are known for both respondents and nonrespondents, and hence cannot directly address the question of nonresponse bias. The approach taken does reflect the reduction in bias obtained through the use of nonresponse weighting adjustments. As such, it is more appropriate than a simple comparison of the characteristics of nonrespondents with those of nonrespondents for each state.

The second approach is to present some summary characteristics of nonrespondents and respondents from nonresponse adjustment classes where relatively large adjustment factors were obtained. In such classes the number of nonrespondents is relatively large, particularly in relation to the number of respondents available, and hence it is in these cases that the greatest potential for nonresponse bias exists. For those states and classes not appearing in these tables, it can be assumed that the potential for nonresponse bias is likely to be much less than in the cases shown.

Table 7-1
Unweighted and Weighted Counts of Assessed Students by State and Grade

State	Grade 8 Mathematics		Grade 4 Mathematics	
	Unweighted	Weighted	Unweighted	Weighted
Alabama	2,522	51,202	2,605	52,211
Arizona	2,617	41,443	2,741	48,978
Arkansas	2,556	30,269	2,621	31,479
California	2,516	314,375	2,412	333,787
Colorado	2,799	40,122	2,906	45,706
Connecticut	2,613	29,459	2,600	31,592
Delaware	1,934	7,000	2,040	65,966
District of Columbia	1,816	4,265	2,399	5,183
Florida	2,549	116,605	2,828	140,142
Georgia	2,589	77,420	2,766	92,994
Guam	1,496	1,667	1,933	2,164
Hawaii	2,454	11,316	2,625	12,687
Idaho	2,615	16,646	2,784	16,925
Indiana	2,659	75,493	2,593	72,237
Iowa	2,816	35,438	2,770	35,450
Kentucky	2,756	45,515	2,703	43,920
Louisiana	2,582	48,803	2,792	56,033
Maine	2,464	15,318	1,898	10,406
Maryland	2,399	47,980	2,844	55,366
Massachusetts	2,456	52,806	2,549	58,076
Michigan	2,616	106,326	2,412	110,356
Minnesota	2,471	49,746	2,640	55,824
Mississippi	2,498	34,209	2,712	37,655
Missouri	2,666	56,203	2,509	53,922
Nebraska	2,285	19,703	2,327	16,538
New Hampshire	2,535	12,129	2,265	14,111
New Jersey	2,174	80,894	2,231	74,579
New Mexico	2,561	20,543	2,342	22,021
New York	2,158	164,133	2,284	180,333
North Carolina	2,769	80,460	2,884	76,824
North Dakota	2,314	8,418	2,193	8,079
Ohio	2,535	136,222	2,637	133,945
Oklahoma	2,141	38,711	2,254	42,298
Pennsylvania	2,612	113,724	2,740	125,005
Rhode Island	2,120	9,621	2,390	10,174
South Carolina	2,625	44,122	2,771	48,038
Tennessee	2,485	57,901	2,708	57,565
Texas	2,614	221,818	2,623	244,988
Utah	2,726	31,181	2,799	34,536
Virgin Islands	1,479	1,601	905	1,863
Virginia	2,710	69,751	2,786	76,029
West Virginia	2,690	23,681	2,786	23,030
Wisconsin	2,814	57,227	2,780	58,196
Wyoming	2,444	7,038	2,605	7,438
TOTAL	108,250	2,508,504	110,992	2,724,651

Table 7-2
Unweighted and Weighted Counts of Excluded Students with Returned Questionnaires by State and Grade

State	Grade 8 Mathematics		Grade 4 Mathematics	
	Unweighted	Weighted	Unweighted	Weighted
Alabama	155	2,952	122	2,468
Arizona	173	2,510	148	2,542
Arkansas	176	2,013	154	1,807
California	237	28,321	327	46,376
Colorado	134	1,809	158	2,526
Connecticut	186	2,077	173	2,340
Delaware	95	303	120	419
District of Columbia	207	454	240	517
Florida	193	8,020	265	13,284
Georgia	135	3,799	150	5,130
Guam	56	72	133	142
Hawaii	138	578	166	814
Idaho	89	545	100	605
Indiana	135	3,623	91	2,483
Iowa	129	1,498	96	1,196
Kentucky	134	2,176	99	1,631
Louisiana	120	2,178	118	2,458
Maine	116	710	116	919
Maryland	119	2,358	117	2,337
Massachusetts	213	4,679	214	4,364
Michigan	183	6,864	130	6,132
Minnesota	88	1,785	93	1,940
Mississippi	203	2,643	142	1,919
Missouri	126	2,630	107	2,658
Nebraska	108	837	117	854
New Hampshire	152	686	98	548
New Jersey	168	5,900	132	4,382
New Mexico	154	1,158	165	1,747
New York	193	15,234	127	10,073
North Carolina	102	2,709	121	3,267
North Dakota	63	207	44	165
Ohio	174	8,833	156	8,606
Oklahoma	184	2,605	199	3,325
Pennsylvania	127	5,153	112	5,184
Rhode Island	119	517	151	627
South Carolina	170	2,771	142	2,429
Tennessee	136	2,978	114	2,603
Texas	205	15,633	232	20,093
Utah	131	1,425	125	1,452
Virgin Islands	77	86	24	48
Virginia	153	3,890	156	4,203
West Virginia	177	1,454	134	1,113
Wisconsin	127	2,610	135	3,215
Wyoming	107	289	98	272
TOTAL	6,367	159,572	6,161	181,213

7.4.1 Weighted Distributions of Schools Before and After School Nonresponse

Tables 7-3 and 7-4 show the mean values of certain school characteristics, both before and after nonresponse. The means are weighted appropriately to reflect whether nonresponse adjustments have been applied (i.e., to respondents only) or not (to the full set of in-scope schools). The variables for which means are presented are the percentage of students in the school who are Black, the percentage who are Hispanic, the median income of the ZIP code area where the school is located, and the "type of locale." All variables were obtained from the sample frame, described in Chapter 3. The type of locale variable has seven possible levels, which are defined in section 3.4.2. Although this variable is not interval-scaled, the mean value does give an indication of the degree of urbanization of the population represented by the school sample (lower values for type of locale indicate a greater degree of urbanization).

Two sets of means are presented for these four variables. The first set shows the weighted mean derived from the full sample of in-scope schools; that is, respondents and nonrespondents (for which there was no participating substitute). The weight for each sampled school is the product of the school base weight and the grade enrollment. This weight therefore represents the number of students in the state represented by the selected school. The second set of means is derived from responding schools only, after school substitution. In this case the weight for each school is the product of the nonresponse-adjusted school weight and the grade enrollment, and therefore indicates the number of students in the state represented by the responding school.

The differences between these sets of means give an indication of the potential for nonresponse bias that has been introduced by nonresponding schools for which there was no participating substitute. For example, in Arkansas at grade 4 the mean percentage Black enrollment, estimated from the original sample, is 24.92 percent. The estimate from the responding schools is 24.47 percent. Thus there may be a slight bias in the results for Arkansas because these two means differ. Note, however, that throughout these two tables the differences in the two sets of mean values are very slight, suggesting that it is unlikely that substantial bias has been introduced by schools that did not participate and for which no substitute participated. Of course in a number of states (as indicated) there was no nonresponse at the school level, so that these sets of means are identical. Even in those states where school nonresponse was relatively high (such as Maine, New Jersey, and New York), the differences in means are slight.

7.4.2 Characteristics of Nonresponding Schools

Tables 7-5 and 7-6 show the distributions of some characteristics of nonresponding and responding schools, by school nonresponse adjustment class, for classes with adjustment factors in excess of 1.25. Table 7-5 shows results for grade 4, Table 7-6 for grade 8. The respondents include the case where substitute schools participated. In other words, the nonrespondents include only those nonrespondents for which no substitute participated.

Table 7-3
Weighted Mean Values Derived from Sampled Schools, Grade 4

State	Weighted Participation Rate After Substitution	Weighted Mean Values Derived from Full Sample				Weighted Mean Values Derived from Responding Sample, with Substitutes and School Nonresponse Adjustment			
		Percent Black	Percent Hispanic	Median Income	Type of Locale	Percent Black	Percent Hispanic	Median Income	Type of Locale
Alabama	97%	31.72%	0.04%	$22,359	4.67	31.46%	0.04%	$22,455	4.67
Arizona	100%	4.06%	21.56%	$29,783	3.17	4.06%	21.56%	$29,783	3.17
Arkansas	99%	24.92%	0.38%	$21,375	5.38	24.47%	0.39%	$21,405	5.38
California	97%	8.32%	35.91%	$32,636	3.17	8.35%	35.89%	$32,682	3.16
Colorado	100%	4.50%	17.28%	$31,517	3.67	4.50%	17.28%	$31,517	3.67
Connecticut	99%	9.82%	8.46%	$39,532	3.66	9.82%	8.46%	$39,560	3.65
Delaware	92%	24.25%	0.32%	$25,543	4.48	23.20%	0.32%	$25,290	4.48
Dist. of Columbia	99%	90.81%	3.47%	$27,916	1.00	91.07%	3.49%	$27,901	1.00
Florida	100%	24.17%	10.98%	$27,580	3.59	24.17%	10.98%	$27,580	3.59
Georgia	100%	33.93%	1.34%	$28,206	4.41	33.93%	1.34%	$28,206	4.41
Guam	94%	2.27%	0.31%	-	7.00	2.37%	0.31%	-	7.00
Hawaii	100%	1.41%	0.00%	$33,990	4.00	1.41%	0.00%	$33,990	4.00
Idaho	97%	0.12%	4.38%	$25,520	5.42	0.12%	4.88%	$25,558	5.42
Indiana	91%	11.36%	0.59%	$28,441	4.35	10.97%	0.58%	$28,595	4.35
Iowa	100%	0.96%	0.25%	$26,228	4.92	0.96%	0.25%	$26,228	4.92
Kentucky	96%	7.37%	0.07%	$22,565	5.28	7.40%	0.06%	$22,561	5.29
Louisiana	100%	45.18%	0.82%	$22,414	4.28	45.18%	0.82%	$22,414	4.28
Maine	71%	0.17%	0.54%	$27,022	5.73	0.80%	0.52%	$26,750	5.72
Maryland	99%	27.89%	1.36%	$39,729	3.45	27.77%	1.38%	$39,949	3.45
Massachusetts	97%	6.77%	4.12%	$37,119	3.71	6.79%	4.12%	$37,123	3.70
Michigan	90%	14.79%	0.96%	$31,784	4.11	14.60%	1.12%	$31,913	4.10
Minnesota	94%	2.03%	0.55%	$32,178	4.73	1.97%	0.54%	$32,426	4.73
Mississippi	100%	47.94%	0.17%	$19,437	5.58	47.94%	0.17%	$19,437	5.58
Missouri	97%	15.40%	0.66%	$27,231	4.47	15.26%	0.64%	$27,080	4.51
Nebraska	87%	3.94%	0.93%	$27,906	4.79	3.83%	1.04%	$27,981	4.80
New Hampshire	80%	0.73%	0.86%	$35,647	5.21	0.65%	0.69%	$35,716	5.22
New Jersey	82%	15.73%	8.37%	$40,236	3.60	14.84%	8.65%	$40,099	3.59
New Mexico	90%	2.60%	44.47%	$22,600	4.63	2.74%	45.67%	$22,855	4.64
New York	83%	15.75%	16.04%	$32,217	3.17	14.70%	16.17%	$32,314	3.18
North Carolina	99%	27.54%	0.01%	$26,066	4.94	27.21%	0.01%	$26,131	4.94
North Dakota	90%	0.46%	0.06%	$26,916	5.07	0.28%	0.07%	$26,624	5.11
Ohio	91%	10.24%	0.35%	$28,765	4.12	9.55%	0.32%	$28,965	4.15
Oklahoma	98%	7.46%	1.32%	$25,421	4.49	6.52%	1.33%	$25,444	4.50
Pennsylvania	95%	12.82%	3.30%	$28,541	4.27	12.96%	3.32%	$28,547	4.26
Rhode Island	96%	4.30%	3.85%	$30,169	3.38	3.90%	4.00%	$30,057	3.36
South Carolina	99%	37.51%	0.07%	$26,503	5.00	37.42%	0.07%	$26,359	5.00
Tennessee	93%	20.88%	0.06%	$24,437	4.12	21.86%	0.06%	$24,526	4.12
Texas	98%	14.45%	34.15%	$26,340	3.44	14.30%	34.54%	$26,272	3.44
Utah	98%	0.13%	0.84%	$31,139	4.25	0.13%	0.84%	$31,149	4.25
Virginia	99%	24.95%	1.38%	$36,618	4.19	24.92%	1.38%	$36,445	4.19
West Virginia	100%	2.80%	0.16%	$21,548	5.63	2.80%	0.16%	$21,548	5.63
Wisconsin	100%	6.80%	1.34%	$31,250	4.37	6.80%	1.34%	$31,250	4.37
Wyoming	97%	0.63%	6.57%	$31,003	5.40	0.64%	6.68%	$30,948	5.40

Table 7-4
Weighted Mean Values Derived from Sampled Schools, Grade 8

State	Weighted Participation Rate After Substitution	Weighted Mean Values Derived from Full Sample				Weighted Mean Values Derived from Responding Sample, with Substitutes and School Nonresponse Adjustment			
		Percent Black	Percent Hispanic	Median Income	Type of Locale	Percent Black	Percent Hispanic	Median Income	Type of Locale
Alabama	92%	31.33%	0.03%	$21,746	4.92	31.98%	0.02%	$21,636	4.92
Arizona	99%	3.74%	22.74%	$29,074	3.30	3.75%	22.59%	$28,987	3.30
Arkansas	97%	24.13%	0.33%	$21,692	5.44	23.47%	0.33%	$21,632	5.44
California	98%	8.89%	31.50%	$34,481	3.19	8.65%	31.32%	$34,641	3.19
Colorado	100%	4.64%	16.51%	$31,468	3.71	4.64%	16.51%	$31,468	3.71
Connecticut	99%	8.94%	6.58%	$40,763	3.92	8.94%	6.60%	$40,787	3.91
Delaware	100%	20.91%	0.61%	$32,463	5.19	20.91%	0.61%	$32,463	5.19
Dist. of Columbia	100%	92.49%	4.08%	$28,443	1.00	92.49%	4.08%	$28,443	1.00
Florida	100%	23.35%	11.58%	$28,009	3.55	23.35%	11.58%	$28,009	3.55
Georgia	99%	34.51%	0.94%	$28,048	4.46	34.5%	0.90%	$27,602	4.47
Guam	100%	1.50%	0.00%	-	7.00	1.50%	0.00%	-	7.00
Hawaii	100%	0.93%	0.08%	$33,823	3.91	0.93%	0.08%	$33,825	3.91
Idaho	91%	0.13%	4.75%	$25,298	5.51	0.13%	4.49%	$25,577	5.51
Indiana	94%	9.16%	0.90%	$28,982	4.62	8.94%	0.42%	$29,046	4.62
Iowa	99%	1.20%	0.27%	$25,962	5.15	1.20%	0.27%	$25,997	5.15
Kentucky	98%	7.09%	0.12%	$22,603	5.13	7.18%	0.13%	$22,602	5.12
Louisiana	100%	44.78%	1.39%	$22,819	4.31	44.78%	1.39%	$22,819	4.31
Maine	84%	0.17%	0.53%	$27,012	5.75	0.09%	0.50%	$27,092	5.75
Maryland	91%	29.57%	1.14%	$40,625	3.53	29.13%	1.13%	$40,213	3.53
Massachusetts	95%	3.63%	5.07%	$37,351	3.90	3.50%	5.35%	$37,302	3.91
Michigan	94%	16.78%	0.94%	$32.144	4.18	16.61%	0.85%	$32,232	4.18
Minnesota	92%	1.59%	0.45%	$31,516	4.74	0.76%	0.43%	$32,022	4.75
Mississippi	100%	44.03%	0.19%	$19,000	5.49	44.03%	0.19%	$19,000	5.49
Missouri	99%	11.49%	0.90%	$26,929	4.77	11.55%	0.90%	$26,944	4.77
Nebraska	85%	2.97%	0.66%	$27,231	5.22	3.25%	0.68%	$26,933	5.13
New Hampshire	92%	0.76%	0.86%	$35,647	5.21	0.65%	0.69%	$35,716	5.22
New Jersey	78%	16.65%	7.87%	$40,552	3.67	15.85%	8.63%	$40,820	3.67
New Mexico	94%	2.03%	43.94%	$22,788	4.65	2.13%	44.30%	$22,967	4.65
New York	83%	17.06%	12.20%	$33,596	3.25	16.17%	11.22%	$34,276	3.27
North Carolina	98%	26.78%	0.07%	$26,064	5.00	26.26%	0.07%	$26,059	5.00
North Dakota	97%	0.19%	0.03%	$26,306	5.28	0.19%	0.03%	$26,306	5.28
Ohio	90%	11.30%	0.21%	$28,658	4.30	11.58%	0.26%	$28,447	4.30
Oklahoma	98%	6.45%	0.72%	$24,635	4.81	6.51%	0.73%	$24,622	4.81
Pennsylvania	94%	10.81%	1.34%	$29,282	4.46	11.32%	1.43%	$29,109	4.43
Rhode Island	100%	3.74%	3.29%	$30,846	3.60	3.74%	3.29%	$30,845	3.60
South Carolina	97%	36.33%	0.24%	$26,141	4.99	36.14%	0.25%	$26,210	4.99
Tennessee	91%	18.93%	0.09%	$23,731	4.39	19.27%	0.10%	$23,785	4.38
Texas	99%	12.92%	31.62%	$26,923	3.52	12.76%	31.89%	$26,872	3.52
Utah	100%	0.10%	0.98%	$30,655	4.22	0.10%	0.98%	$30,655	4.22
Virginia	97%	22.50%	1.66%	$36,906	4.25	22.53%	1.69%	$37,021	4.25
Virgin Islands	100%	87.96%	10.01%	-	-	87.96%	10.01%	-	-
West Virginia	100%	3.88%	0.12%	$21,623	5.60	3.88%	0.12%	$21,623	5.60
Wisconsin	100%	5.31%	0.98%	$31,446	4.76	5.31%	0.98%	$31,446	4.76
Wyoming	99%	0.63%	6.69%	$30,977	5.47	0.63%	6.74%	$30,962	5.47

Table 7-5
Grade 4 School Nonresponse Adjustment Classes with Adjustment Factors Greater than 1.25

State	Class	Nonresponse Adjustment	Response Status	Number of School Selections[1]	Types of Locale	Percent of State Student Population Represented	Enrollment Percent Black			Enrollment Percent Hispanic			Median Household Income ($)[2]		
							Min.	Med.	Max.	Min.	Med.	Max.	Min.	Med.	Max.
Delaware	10	1.30093	Respondents	18	7	26.08%	0%	20%	50%	0%	0%	0%	20,300	24,158	30,976
			Nonrespondents	6	7	7.85%	10%	40%	40%	0%	0%	0%	23,655	28,723	28,723
Guam	2	1.28151	Respondents	6	7	20.67%	0%	1%	8%	0%	0%	0%	–	–	–
			Nonrespondents	1	7	5.82%	0%	–	–	0%	–	–	–	–	–
Massachusetts	7	1.262	Respondents	11	5,6,7	7.92%	0%	0%	10%	0%	0%	0%	24,140	30,706	33,931
			Nonrespondents	3	6,7	2.07%	0%	0%	0%	0%	0%	0%	25,519	28,655	31,577
Maine	3	1.39336	Respondents	14	6	12.09%	0%	0%	0%	0%	0%	1%	20,263	22,892	24,132
			Nonrespondents	7	6	4.76%	0%	0%	8%	0%	0%	3%	21,226	22,580	24,498
	4	1.77427	Respondents	11	6	8.24%	0%	0%	3%	0%	0%	4%	24,506	25,664	28,452
			Nonrespondents	9	6	6.38%	0%	0%	0%	0%	0%	2%	25,747	27,149	28,452
	5	1.55119	Respondents	13	6	11.18%	0%	0%	1%	0%	1%	2%	28,552	30,357	40,726
			Nonrespondents	7	6	6.16%	0%	0%	1%	0%	0%	2%	28,531	31,225	50,455
	6	1.74673	Respondents	11	7	4.69%	0%	0%	0%	0%	0%	1%	15,148	20,598	22,559
			Nonrespondents	9	7	3.51%	0%	0%	0%	0%	0%	1%	18,684	20,735	22,675
	7	1.30330	Respondents	14	7	9.31%	0%	0%	0%	0%	0%	1%	22,965	24,114	25,237
			Nonrespondents	5	7	2.82%	0%	0%	0%	0%	0%	1%	23,708	24,486	25,356
	8	1.30097	Respondents	14	7	9.40%	0%	0%	0%	0%	0%	1%	25,600	28,073	38,719
			Nonrespondents	5	7	2.83%	0%	0%	0%	0%	0%	1%	26,507	29,222	45,690
Michigan	1	1.28032	Respondents	7	2	6.10%	0%	10%	20%	0%	0%	0%	26,546	28,607	40,794
			Nonrespondents	2	2	1.71%	10%	–	20%	0%	–	0%	22,972	–	30,675
	8	1.25481	Respondents	8	5,6	6.84%	0%	0%	0%	0%	0%	1%	26,602	29,196	32,740
			Nonrespondents	2	6	1.74%	0%	–	0%	0%	–	0%	27,429	–	31,422
Minnesota	1	1.26531	Respondents	10	1,2	9.61%	0%	0%	30%	0%	0%	1%	22,370	32,283	47,213
			Nonrespondents	3	1,2	2.55%	0%	0%	31%	0%	0%	2%	16,845	20,316	35,877

---

[1] In some states, larger schools were selected into the sample more than once. Thus, the number of school selections may exceed somewhat the actual number of schools involved.

[2] Median household income of ZIP code area where school is located, derived from 1980 population census data and expressed in 1985 dollars. These data are not available for Guam.

Table 7-5 (continued)
Grade 4 School Nonresponse Adjustment Classes with Adjustment Factors Greater than 1.25

State	Class	Nonresponse Adjustment	Response Status	Number of School Selections[1]	Types of Locale	Percent of State Student Population Represented	Enrollment Percent Black			Enrollment Percent Hispanic			Median Household Income ($)[2]		
							Min.	Med.	Max.	Min.	Med.	Max.	Min.	Med.	Max.
Nebraska	5	1.31472	Respondents	10	6	7.39%	0%	0%	0%	0%	0%	20%	26,105	27,263	27,724
			Nonrespondents	5	5,6	2.33%	0%	0%	0%	0%	0%	1%	25,877	26,148	27,724
	8	1.28562	Respondents	27	7	18.34%	0%	0%	0%	0%	0%	10%	21,189	25,590	42,293
			Nonrespondents	14	7	5.24%	0%	0%	0%	0%	0%	10%	22,191	25,532	33,080
New Hampshire	2	1.28341	Respondents	7	2,4	7.13%	0%	1%	3%	0%	0%	3%	33,697	36,315	38,374
			Nonrespondents	2	2	2.02%	3%	–	3%	3%	–	16%	33,067	–	33,067
	4	1.26630	Respondents	31	5,6	28.14%	0%	1%	4%	0%	0%	2%	21,480	32,235	39,464
			Nonrespondents	8	6	7.50%	0%	0.5%	1%	0%	0%	1%	25,044	28,524	33,494
New Jersey	5	1.40051	Respondents	14	5,6	12.77%	0%	0%	1%	0%	0%	2%	39,457	46,309	56,506
			Nonrespondents	5	6	5.11%	0%	1%	1%	0%	1%	1%	41,444	46,202	63,873
	5	1.26380	Respondents	37	3	31.88%	0%	0%	95%	0%	0%	50%	21,861	41,674	72,889
			Nonrespondents	8	3	8.41%	0%	0%	80%	0%	0%	10%	32,153	42,112	75,500
	6	1.37191	Respondents	17	4	14.51%	0%	0%	30%	0%	0%	72%	23,088	43,645	72,339
			Nonrespondents	7	4	5.40%	0%	0%	50%	0%	0%	0%	34,731	41,828	49,195
	8	1.37400	Respondents	13	5,6,7	10.15%	0%	0%	40%	0%	0%	2%	34,970	45,526	68,749
			Nonrespondents	5	6,7	3.80%	0%	0%	0%	0%	0%	0%	37,710	45,587	56,951
New Mexico	7	1.38274	Respondents	7	6	7.29%	0%	0%	17%	4%	11%	25%	15,417	23,665	44,495
			Nonrespondents	3	6	2.79%	0%	0%	1%	0%	12%	31%	13,792	21,755	24,393
	10	1.37083	Respondents	17	7	11.34%	0%	0%	2%	0%	62%	99%	10,393	16,865	25,713
			Nonrespondents	6	7	4.21%	0%	0%	3%	0%	34.5%	94%	13,988	16,076	24,493
New York	2	1.35294	Respondents	17	1,2	16.52%	1%	21%	57%	25%	55%	97%	13,243	18,701	35,674
			Nonrespondents	6	1,2	5.83%	3%	35%	79%	20%	54%	91%	12,271	15,599	24,336
	3	1.32967	Respondents	9	1,2	8.85%	32%	45%	94%	0%	2%	13%	16,589	22,304	39,488
			Nonrespondents	3	1	2.92%	30%	92%	94%	5%	8%	17%	11,411	15,319	31,124

---

[1] In some states, larger schools were selected into the sample more than once. Thus, the number of school selections may exceed somewhat the actual number of schools involved.

[2] Median household income of ZIP code area where school is located, derived from 1980 population census data and expressed in 1985 dollars.

BEST COPY AVAILABLE

Table 7-5 (continued)
Grade 4 School Nonresponse Adjustment Classes with Adjustment Factors Greater than 1.25

State	Class	Nonresponse Adjustment	Response Status	Number of School Selections[1]	Types of Locale	Percent of State Student Population Represented	Enrollment Percent Black			Enrollment Percent Hispanic			Median Household Income ($)[2]		
							Min.	Med.	Max.	Min.	Med.	Max.	Min.	Med.	Max.
New York	9	1.28169	Respondents	7	6,7	6.90%	0%	0%	0%	0%	0%	0%	32,603	38,498	75,500
			Nonrespondents	2	6	1.94%	0%	–	0%	0%	–	0%	40,013	–	46,366
Ohio	9	1.34462	Respondents	6	6	4.83%	0%	0%	10%	0%	0%	20%	27,323	29,847	36,663
			Nonrespondents	2	5,6	1.66%	0%	–	0%	0%	–	10%	28,550	–	28,972
	10	1.30435	Respondents	7	7	5.47%	0%	0%	0%	0%	0%	0%	13,471	21,213	23,580
			Nonrespondents	3	7	1.66%	0%	0%	0%	0%	0%	0%	15,664	17,135	23,078
Tennessee	8	1.28668	Respondents	8	5,6	6.80%	0%	5%	45%	0%	0%	1%	19,549	19,997	21,905
			Nonrespondents	2	6	1.95%	0%	–	0%	0%	–	0%	20,319	–	21,111

---

[1] In some states, larger schools were selected into the sample more than once. Thus, the number of school selections may exceed somewhat the actual number of schools involved.

[2] Median household income of ZIP code area where school is located, derived from 1980 population census data and expressed in 1985 dollars.

Table 7-6
Grade 8 School Nonresponse Adjustment Classes with Adjustment Factors Greater than 1.25

State	Class	Nonresponse Adjustment	Response Status	Number of School Selections[1]	Types of Locale	Percent of State Student Population Represented	Enrollment Percent Black			Enrollment Percent Hispanic			Median Household Income ($)[2]		
							Min.	Med.	Max.	Min.	Med.	Max.	Min.	Med.	Max.
Alabama	12	1.30476	Respondents	7	7	6.64%	30%	60%	90%	0%	0%	0%	9,999	16,234	22,382
			Nonrespondents	2	7	2.02%	40%	–	70%	0%	–	0%	11,951	–	13,192
Arkansas	9	1.27778	Respondents	9	7	8.74%	22%	34%	62%	0%	0%	2%	13,009	20,138	48,100
			Nonrespondents	3	7	2.43%	26%	91%	92%	0%	0%	1%	16,845	17,308	42,921
Idaho	4	1.46154	Respondents	13	5	12.11%	0%	0%	0%	1%	4%	6%	26,711	28,714	29,477
			Nonrespondents	6	5	5.59%	0%	0%	0%	6%	13%	13%	21,391	24,804	26,711
Maryland	7	1.36364	Respondents	11	5,6,7	10.59%	8%	20%	88%	0%	0%	1%	23,460	31,414	60,213
			Nonrespondents	4	6,7	3.85%	11%	25%	44%	0%	0.5%	1%	36,252	43,980	54,956
Maine	3	1.27944	Respondents	15	6	13.72%	0%	0%	0%	0%	0%	1%	20,158	22,722	24,506
			Nonrespondents	4	6	3.83%	0%	0%	8%	0%	1%	3%	20,464	22,229	23,482
	6	1.38466	Respondents	17	7	14.49%	0%	0%	1%	0%	0%	1%	18,047	21,912	26,400
			Nonrespondents	9	7	5.57%	0%	0%	0%	0%	0%	0%	12,776	21,350	24,614
Michigan	11	1.25641	Respondents	8	7	7.28%	0%	0%	10%	0%	0%	0%	23,547	29,037	47,882
			Nonrespondents	2	7	1.87%	0%	–	20%	0%	–	0%	27,593	–	29,395
Minnesota	1	1.41546	Respondents	9	1,2	9.43%	0%	0%	17%	0%	0%	2%	24,771	31,563	58,266
			Nonrespondents	4	1,2	3.92%	0%	28%	38%	0%	1%	1%	16,726	23,045	26,661
Nebraska	1	1.26316	Respondents	19	2,4	17.91%	0%	6%	30%	0%	0%	2%	14,051	28,015	34,530
			Nonrespondents	5	4	4.71%	0%	0%	10%	0%	0%	0%	31,757	34,522	34,522
	4	1.26189	Respondents	7	6	6.60%	0%	0%	0%	0%	0%	0%	26,305	27,063	27,769
			Nonrespondents	4	5,6	1.73%	0%	0%	0%	0%	0%	0%	26,148	27,724	27,724
	7	1.28035	Respondents	13	7	10.17%	0%	0%	0%	0%	0%	10%	20,640	22,793	24,766
			Nonrespondents	6	7	2.85%	0%	0%	0%	0%	0%	15%	20,693	22,479	23,004
New Hampshire	2	1.28271	Respondents	7	2	6.64%	1%	2%	3%	0%	2%	4%	33,067	33,067	33,697
			Nonrespondents	2	2	1.88%	3%	–	3%	7%	–	7%	33,067	–	33,067

---

[1] In some states, larger schools were selected into the sample more than once. Thus, the number of school selections may exceed somewhat the actual number of schools involved.

[2] Median household income of ZIP code area where school is located, derived from 1980 population census data and expressed in 1985 dollars.

Table 7-6 (continued)
Grade 8 School Nonresponse Adjustment Classes with Adjustment Factors Greater than 1.25

State	Class	Nonresponse Adjustment	Response Status	Number of School Selections[1]	Types of Locale	Percent of State Student Population Represented	Enrollment Percent Black			Enrollment Percent Hispanic			Median Household Income ($)[2]		
							Min.	Med.	Max.	Min.	Med.	Max.	Min.	Med.	Max.
New Jersey	3	1.28067	Respondents	32	3	29.76%	0%	0%	98%	0%	0%	83%	22,737	44,720	74,026
			Nonrespondents	7	3	8.35%	0%	0%	60%	0%	0%	0%	32,153	39,244	75,500
	4	1.33333	Respondents	6	4	5.57%	0%	0%	90%	0%	15%	80%	23,088	30,663	35,399
			Nonrespondents	2	4	1.86%	12%	–	83%	2%	–	9%	31,427	–	33,097
	6	1.33333	Respondents	6	4	5.57%	0%	0%	20%	0%	0%	0%	49,616	58,348	69,221
			Nonrespondents	2	4	1.86%	0%	–	0%	0%	–	2%	47,598	–	55,652
	7	1.51173	Respondents	16	5,6,7	14.51%	0%	0%	0%	0%	0%	0%	26,390	38,332	65,917
			Nonrespondents	8	6,7	7.42%	0%	0%	60%	0%	0%	0%	22,539	42,143	75,500
New Mexico	6	1.33333	Respondents	9	6	8.41%	0%	1%	5%	9%	26%	33%	17,529	22,243	44,495
			Nonrespondents	3	6	2.80%	0%	0%	0%	0%	23%	25%	13,792	23,235	24,393
New York	1	1.55556	Respondents	27	1,2	26.23%	0%	27%	96%	0%	16%	88%	12,475	25,091	67,635
			Nonrespondents	15	1	14.57%	2%	24%	93%	6%	24%	83%	12,457	21,893	48,053
Ohio	7	1.33333	Respondents	6	5,6	5.26%	0%	0%	10%	0%	0%	0%	17,711	21,471	24,885
			Nonrespondents	2	6	1.75%	0%	–	0%	0%	–	0%	23,071	–	24,258
Tennessee	11	1.33333	Respondents	6	7	5.79%	0%	0%	60%	0%	0%	0%	16,783	17,627	19,216
			Nonrespondents	2	7	1.93%	0%	–	0%	0%	–	0%	16,420	–	18,405

---

[1] In some states, larger schools were selected into the sample more than once. Thus, the number of school selections may exceed somewhat the actual number of schools involved.

[2] Median household income of ZIP code area where school is located, derived from 1980 population census data and expressed in 1985 dollars.

The characteristics shown are as follows:

- *The set of distinct values for the "type of locale" variable.* This variable, which was used for sample stratification, has seven possible levels, which are defined in Chapter 3, section 3.4.2.

- *The percentage of the state's public-school grade enrollment represented in the sample by the schools within the adjustment class.* The school nonresponse adjustment factor is calculated directly from these two quantities (one for respondents, one for nonrespondents). The potential for nonresponse bias is generally greater in cases where the size of the set of nonrespondents is relatively large.

- *The minimum, median, and maximum percentage enrollments of Black and Hispanic students.* In cases where there are only two nonresponding school/hits involved, only the minimum and maximum are presented. In the case of a single nonresponding school, the value for that school is presented as the minimum.

- *The minimum, median, and maximum household incomes of the five digit ZIP code area where the school is located.* The data are calculated from 1980 Census data, but are updated to 1985 dollars.

Examination of the table shows that invariably the respondents and nonrespondents are quite similar with regard to type of locale. There are great similarities in many cases for other characteristics also, but on some occasions the nonresponding schools have a somewhat lower median income distribution than the respondents, and occasionally also there is some difference in the distributions of minority enrollment levels. For example, for grade 4, in New York, Class 3, the nonresponding schools have somewhat higher rates of Black and Hispanic enrollment and somewhat lower median household incomes than the respondents. For grade 8, in Minnesota, Class 1, the nonresponding schools have somewhat greater Hispanic enrollment and noticeably lower median income than the respondents. By contrast, in Nebraska, Class 1, the nonresponding schools have somewhat lower Hispanic enrollment and noticeably higher median income than the respondents.

7.4.3 Weighted Distributions of Students Before and After Student Absenteeism

Tables 7-7 and 7-8 show, for each state, the weighted sampled percentages of students by gender (male) and race/ethnicity (White, not Hispanic; Black, not Hispanic; Hispanic) for the full sample of students (after student exclusion) and for the assessed sample. Table 7-7 shows results for grade 4; Table 7-8 shows results for grade 8.

The weight used for the full sample is the adjusted student base weight, defined in section 7.3.5. The weight for the assessed students is the final student weight, also defined section 7.3.5. The difference between the estimates of the population subgroups is an estimate of the bias in estimating the size of the subgroup, resulting from student absenteeism from the assessment. As such it is an indicator of the potential for nonresponse bias in the assessment results, resulting from student absenteeism.

Table 7-7
Weighted Student Percentages Derived from Sampled Schools, Grade 4

State	Weighted Student Participation	Weighted Percentages Derived from Full Sample				Weighted Percentages Derived from Assessed Sample, with Student Nonresponse Adjustment			
		Percent Male	Percent White	Percent Black	Percent Hispanic	Percent Male	Percent White	Percent Black	Percent Hispanic
Arizona	95%	51.25	56.17	4.10	28.48	51.35	55.96	4.17	28.70
Arkansas	96%	52.92	69.68	21.03	6.09	53.06	69.21	21.10	6.40
California	94%	51.14	44.78	6.50	35.34	51.61	44.77	6.43	34.97
Colorado	95%	49.78	67.79	5.33	21.48	49.93	67.53	5.22	21.85
Connecticut	96%	48.62	72.94	10.22	13.50	48.71	72.89	10.18	13.49
Delaware	95%	51.27	65.74	22.99	7.91	50.63	65.87	22.73	8.04
Dist. of Columbia	93%	48.12	5.30	82.42	9.16	47.75	5.34	82.10	9.52
Florida	95%	48.44	58.16	21.34	17.09	48.37	57.90	21.28	17.29
Georgia	95%	51.76	55.97	35.56	5.93	51.46	55.67	35.49	6.19
Guam	95%	51.47	12.26	3.87	19.08	51.64	12.07	3.81	19.80
Hawaii	95%	49.85	20.81	4.14	17.18	49.24	20.70	4.27	17.90
Idaho	97%	49.59	84.14	0.57	11.22	49.18	83.86	0.58	11.45
Indiana	96%	49.79	81.99	10.35	5.19	50.30	82.08	9.95	5.38
Iowa	96%	50.99	89.56	2.34	5.14	50.83	89.53	2.22	5.29
Kentucky	96%	49.58	84.92	8.74	4.04	49.47	84.80	8.63	4.29
Louisiana	95%	51.90	49.56	42.82	4.78	51.77	49.50	42.54	5.03
Maine	95%	49.41	91.21	0.61	4.92	49.02	90.93	0.59	5.75
Maryland	96%	49.18	58.77	30.13	5.79	49.63	58.52	30.01	5.96
Massachusetts	95%	50.65	78.77	7.76	7.96	50.83	79.15	7.46	8.04
Michigan	94%	52.04	73.43	13.95	8.27	51.71	73.34	13.35	8.74
Minnesota	95%	50.23	85.79	2.64	6.90	50.26	85.44	2.59	7.29
Mississippi	97%	50.10	40.76	51.89	5.60	51.68	40.40	52.06	5.83
Missouri	96%	51.77	76.96	14.26	5.56	51.95	76.98	13.81	5.86
Nebraska	96%	50.44	83.56	6.73	6.70	50.76	84.12	5.96	7.04
New Hampshire	96%	50.11	89.24	1.22	4.99	50.28	88.92	1.24	5.16
New Jersey	96%	50.89	63.25	15.56	15.02	50.97	65.77	13.69	14.38
New Mexico	95%	47.09	44.53	3.52	46.27	46.71	43.93	3.60	46.65
New York	96%	51.61	61.41	12.50	20.18	51.65	59.31	12.87	21.69
North Carolina	95%	51.20	62.37	28.28	5.30	51.33	61.52	28.81	5.59
North Dakota	96%	52.69	91.29	0.45	3.60	52.93	91.13	0.48	3.82
Ohio	95%	51.23	78.96	11.21	5.81	51.28	79.13	10.67	6.15
Oklahoma	84%	50.51	72.96	9.28	6.62	50.72	72.89	9.18	7.09
Pennsylvania	96%	52.92	76.55	12.65	7.49	52.95	76.94	12.19	7.44
Rhode Island	95%	51.43	77.80	6.31	10.82	51.05	77.77	6.27	10.86
South Carolina	97%	50.28	55.32	37.20	5.50	49.73	54.90	37.33	5.72
Tennessee	96%	52.28	69.78	22.57	5.02	52.49	69.06	23.04	5.25
Texas	96%	49.27	48.70	14.19	33.79	49.12	48.92	14.11	33.52
Utah	96%	50.27	85.97	0.95	9.56	50.71	85.73	0.99	9.73
Virginia	95%	50.84	67.24	23.45	4.69	50.90	67.28	23.16	4.78
West Virginia	96%	49.03	90.40	2.54	4.52	48.82	90.18	2.52	4.73
Wisconsin	96%	51.36	81.11	6.22	7.16	51.47	80.91	6.18	7.32
Wyoming	96%	50.45	82.59	0.93	11.05	50.47	82.41	0.94	11.32

BEST COPY AVAILABLE

Table 7-8
Weighted Student Percentages Derived from Sampled Schools, Grade 8

State	Weighted Student Participation	Weighted Percentages Derived from Full Sample				Weighted Percentages Derived from Assessed Sample, with Student Nonresponse Adjustment			
		Percent Male	Percent White	Percent Black	Percent Hispanic	Percent Male	Percent White	Percent Black	Percent Hispanic
Arizona	93%	51.23	60.70	4.04	27.06	50.83	60.30	3.94	27.61
Arkansas	94%	50.94	72.68	21.55	3.90	50.76	72.35	21.59	4.22
California	92%	49.82	44.18	7.51	35.94	48.92	44.31	6.99	36.07
Colorado	93%	50.47	73.11	4.56	18.06	51.07	73.64	4.17	17.92
Connecticut	94%	50.22	72.13	12.05	12.50	50.09	72.39	11.91	12.26
Delaware	92%	50.41	65.32	25.52	5.68	50.12	64.88	25.47	6.23
D. of Columbia	85%	49.37	2.76	85.83	8.32	48.74	2.82	84.95	9.71
Florida	91%	49.01	56.33	23.02	17.56	48.97	55.51	22.91	18.31
Georgia	93%	48.36	59.18	34.72	2.27	48.03	58.63	34.96	4.04
Guam	90%	52.88	4.85	1.25	13.55	52.13	4.50	1.35	15.19
Hawaii	90%	52.31	17.40	2.84	16.79	51.85	17.18	2.78	18.25
Idaho	95%	51.30	88.19	0.69	7.30	51.26	87.85	0.70	7.49
Indiana	94%	50.61	84.49	8.68	4.39	50.67	84.64	8.45	4.46
Iowa	95%	52.49	92.33	1.87	3.41	52.34	92.34	1.79	3.56
Kentucky	96%	50.03	86.66	8.75	2.86	50.11	86.52	8.78	3.00
Louisiana	93%	46.97	53.95	39.60	4.18	47.17	53.91	39.31	4.57
Maine	92%	51.00	94.04	0.50	1.81	51.11	93.80	0.39	1.96
Maryland	93%	50.38	59.45	30.08	6.13	50.13	59.82	29.40	6.34
Massachusetts	94%	50.23	84.79	4.76	7.69	50.19	83.26	5.35	8.44
Michigan	94%	48.05	72.16	19.58	4.70	48.16	73.26	18.41	4.85
Minnesota	94%	49.78	91.53	1.48	3.34	49.39	91.44	1.56	3.49
Mississippi	95%	48.21	49.38	43.90	5.43	47.87	49.11	43.76	5.78
Missouri	95%	51.95	82.54	11.93	2.82	51.98	82.42	11.86	2.99
Nebraska	96%	53.02	86.72	4.78	5.22	52.67	86.54	4.93	5.52
New Hampshire	94%	50.12	91.82	0.86	2.76	50.37	91.41	0.93	2.85
New Jersey	94%	49.48	59.57	18.60	14.94	49.15	61.12	17.42	14.35
New Mexico	93%	49.79	43.27	2.36	49.16	49.74	43.74	2.38	48.54
New York	92%	48.88	66.22	14.53	12.55	49.12	61.42	17.38	14.29
North Carolina	94%	49.85	68.52	26.61	2.56	49.82	68.36	26.58	2.68
North Dakota	96%	50.97	92.90	0.44	2.54	51.07	93.32	0.41	2.64
Ohio	93%	50.31	79.86	13.78	3.62	50.31	79.74	13.63	3.97
Oklahoma	80%	50.76	73.50	8.05	5.41	49.96	74.54	7.69	6.06
Pennsylvania	94%	50.28	83.57	10.37	3.21	50.39	83.08	10.68	3.26
Rhode Island	93%	49.77	81.10	5.81	7.95	49.95	80.58	5.96	8.15
South Carolina	94%	50.35	58.05	34.91	5.16	50.47	57.89	34.57	5.56
Tennessee	94%	49.53	76.03	20.21	2.39	49.69	75.33	20.61	2.64
Texas	94%	49.09	48.31	12.06	35.66	49.24	47.91	11.95	36.05
Utah	94%	51.65	89.54	0.60	6.56	51.56	89.50	0.62	6.59
Virginia	94%	50.61	69.22	22.00	4.57	50.28	68.96	21.91	4.63
Virgin Islands	92%	54.40	1.18	77.95	20.17	52.74	1.30	76.83	21.12
West Virginia	94%	48.94	91.19	4.27	2.38	48.78	90.80	4.36	2.55
Wisconsin	94%	50.57	85.62	6.87	4.34	50.53	85.85	6.80	4.39
Wyoming	95%	50.12	86.14	0.76	8.60	50.02	86.19	0.82	8.55

Care must be taken in interpreting these results, however. First, note that there is generally very little difference in the proportions estimated from the full sample and those estimated from the assessed students. While this is encouraging, it does not eliminate the possibility that bias exists, either within the state as a whole, or for results for gender and race/ethnicity subgroups, or for other subgroups. Second, on the other hand, where differences do exist they cannot be used to indicate the likely magnitude or direction of the bias with any reliability. For example, at grade 4 in New Jersey, the percentages of Black and Hispanic students in the full sample are respectively 15.57 and 15.21 percent. For assessed students, these percentages are 13.69 for Black students and 14.39 for Hispanic students. While these differences raise the possibility that some bias exists, it is not appropriate to speculate on the magnitude of this bias by considering the assessment results for Black and Hispanic student students, in comparison to other students in the state. This is because the underrepresented Black and Hispanic students may not be typical of students that were included in the sample, and similarly those students within the same racial/ethnic groups who are disproportionately overrepresented may not be typical either. This is because not all students within the same race/ethnicity group receive the same student nonresponse adjustment. Some insight as to the kinds of students who are receiving relatively large adjustments, and the kinds of students that they are being adjusted to represent, are given in the next section. Small sample sizes within nonresponse adjustment classes make this information difficult to interpret, however. One other feature to note is that, for assessed students, information as to the student's gender and race/ethnicity is provided by the student, while for absent students this information is provided by the school. Evidence from past NAEP assessments (see, for example, Rust & Johnson, 1992) indicates that there can be substantial discrepancies between those two sources, especially with regard to classifying students as Hispanic at grade 4.

7.4.4 Characteristics of Absent Students

Tables 7-9 and 7-10 show some characteristics of assessed (responding) and absent (nonresponding) students, by student nonresponse adjustment class, for classes with adjustment factors in excess of 1.25. Table 7-9 shows results for grade 4, Table 7-10 for grade 8.

In addition to information characterizing the class in terms of age class, monitor status, and type of location, the distributions of certain characteristics of assessed and absent students within each class are presented. The characteristics shown are:

- *The percentage of the state's public-school grade enrollment represented in the sample by the students within the adjustment class.* This is given by the sum of the adjusted student base weights ($W_{hij}^{(2)}$, see section 7.3.5) for the responding and nonresponding selected students respectively, within the student-level nonresponse adjustment class. The student nonresponse adjustment factor is calculated directly from these two quantities (one for respondents, one for nonrespondents). The potential for nonresponse bias is generally greater in cases where the size of the population represented by the nonrespondents is relatively large.

Table 7-9
Grade 4 Student Nonresponse Adjustment Classes with Adjustment Factors Greater than 1.25

State	Class	Nonresponse Adjustment	Age Class*	Monitor Status	Types of Locale	Response Status	Percent of State Population	Percent Male	Percent Race/Ethnicity			
									White	Black	Hispanic	Other
Michigan	37	1.25969	2	Monitored	7	Respondents	2.32%	46.3%	84.5%	0.0%	10.2%	5.3%
						Nonrespondents	0.60%	53.9%	100%	0.0%	0.0%	0.0%
Oklahoma	1	1.27065	1	Unmonitored	1,2	Respondents	2.25%	59.9%	83.0%	7.2%	8.4%	1.3%
						Nonrespondents	0.61%	43.0%	77.9%	0.0%	11.3%	10.7%
	3	1.43672	1	Unmonitored	1,2	Respondents	1.77%	65.9%	27.5%	43.7%	19.5%	9.4%
						Nonrespondents	0.77%	66.2%	35.1%	37.2%	18.6%	9.2%
	4	1.54289	1	Monitored	1,2	Respondents	1.90%	54.5%	41.9%	26.2%	10.7%	21.2%
						Nonrespondents	1.03%	39.1%	42.2%	27.8%	10.4%	19.6%
	6	1.29015	1	Both	3,4	Respondents	1.45%	72.9%	91.6%	0.0%	5.6%	2.8%
						Nonrespondents	0.42%	63.2%	81.8%	9.1%	0.0%	9.1%
	7	1.48809	1	Both	3,4	Respondents	0.79%	56.6%	92.1%	3.9%	0.0%	4.0%
						Nonrespondents	0.39%	76.3%	83.9%	0.0%	7.9%	8.2%
	8	1.25579	1	Unmonitored	6	Respondents	1.24%	68.0%	55.0%	5.9%	2.9%	36.2%
						Nonrespondents	0.32%	64.9%	78.1%	0.0%	0.0%	21.9%
	15	1.26790	1	Monitored	7	Respondents	0.95%	57.7%	61.5%	7.7%	11.7%	19.2%
						Nonrespondents	0.25%	71.2%	42.3%	14.5%	0.0%	43.1%
	19	1.26989	1	Monitored	7	Respondents	0.80%	43.2%	96.4%	0.0%	0.0%	3.6%
						Nonrespondents	0.22%	54.0%	100%	0.0%	0.0%	0.0%
	23	1.35913	2	Monitored	1,2	Respondents	2.60%	37.9%	50.3%	37.1%	4.8%	7.8%
						Nonrespondents	0.96%	48.7%	44.3%	42.3%	9.9%	3.5%

---

* Age class 1 consists of students born in September 1981 or earlier. All other students are in age class 2.

Table 7-10
Grade 8 Student Nonresponse Adjustment Classes with Adjustment Factors Greater than 1.25

State	Class	Nonresponse Adjustment	Age Class*	Monitor Status	Types of Locale	Response Status	Percent of State Population	Percent Male	Percent Race/Ethnicity			
									White	Black	Hispanic	Other
Colorado	5	1.35637	1	Unmonitored	1	Respondents	0.62%	67.7%	26.6%	8.9%	55.1%	9.4%
						Nonrespondents	0.22%	60.7%	0.0%	11.9%	88.1%	0.0%
Connecticut	6	1.32657	1	Both	2	Respondents	1.27%	64.5%	35.3%	22.6%	42.0%	0.0%
						Nonrespondents	0.41%	58.4%	25.1%	28.7%	46.2%	0.0%
District of Columbia	1	1.48133	1	Unmonitored	1	Respondents	2.54%	68.7%	2.3%	68.2%	25.8%	3.7%
						Nonrespondents	1.22%	54.1%	0.0%	90.5%	9.5%	0.0%
	2	1.33259	1	Monitored	1	Respondents	5.17%	66.3%	6.4%	71.1%	20.3%	2.3%
						Nonrespondents	1.72%	72.6%	0.0%	89.3%	10.7%	0.0%
	3	1.40297	1	Unmonitored	1	Respondents	6.69%	61.6%	0.4%	80.4%	14.4%	4.8%
						Nonrespondents	2.70%	64.3%	0.0%	100%	0.0%	0.0%
	4	1.43987	1	Monitored	1	Respondents	6.44%	55.0%	0.3%	86.9%	9.8%	3.0%
						Nonrespondents	2.83%	72.4%	0.0%	100%	0.0%	0.0%
Delaware	1	1.36398	1	Both	2	Respondents	1.75%	70.8%	56.0%	33.0%	7.4%	3.6%
						Nonrespondents	0.64%	58.9%	49.7%	50.3%	0.0%	0.0%
	9	1.26048	1	Unmonitored	7	Respondents	1.54%	52.5%	44.4%	36.8%	18.8%	0.0%
						Nonrespondents	0.40%	40.4%	79.3%	20.7%	0.0%	0.0%
Florida	6	1.35132	1	Unmonitored	2	Respondents	1.32%	54.6%	20.7%	55.6%	20.8%	2.9%
						Nonrespondents	0.46%	79.4%	33.5%	66.5%	0.0%	0.0%
	9	1.25248	1	Both	3,4	Respondents	1.56%	55.3%	50.7%	29.9%	12.8%	6.5%
						Nonrespondents	0.39%	33.9%	64.6%	26.0%	9.4%	0.0%
	10	1.25690	1	Both	4	Respondents	1.94%	54.1%	48.7%	33.7%	5.2%	12.3%
						Nonrespondents	0.50%	58.0%	56.1%	37.0%	6.8%	0.0%
	14	1.29497	1	Both	6,7	Respondents	0.74%	54.7%	66.9%	13.0%	20.1%	0.0%
						Nonrespondents	0.22%	89.8%	74.6%	15.2%	10.2%	0.0%

---

* Age class 1 consists of students born in September 1977 or earlier. All other students are in age class 2.

Table 7-10 (continued)
Grade 8 Student Nonresponse Adjustment Classes with Adjustment Factors Greater than 1.25

State	Class	Nonresponse Adjustment	Age Class*	Monitor Status	Types of Locale	Response Status	Percent of State Population	Percent Male	Percent Race/Ethnicity			
									White	Black	Hispanic	Other
Georgia	11	1.34955	1	Unmonitored	3,4	Respondents Nonrespondents	1.57% 0.55%	62.6% 66.4%	27.7% 46.8%	55.3% 53.2%	9.9% 0.0%	7.1% 0.0%
Guam	1	1.31507	1	Both	7	Respondents Nonrespondents	4.38% 1.38%	60.3% 47.8%	2.7% 8.7%	2.7% 0.0%	23.3% 0.0%	71.2% 91.3%
Hawaii	4	1.42306	1	Unmonitored	4	Respondents Nonrespondents	0.73% 0.31%	64.0% 43.2%	13.9% 22.2%	0.0% 0.0%	36.7% 22.2%	49.4% 55.5%
	5	1.34652	1	Unmonitored	4	Respondents Nonrespondents	1.29% 0.45%	57.2% 77.3%	17.1% 20.0%	6.6% 6.6%	21.0% 0.0%	55.3% 73.4%
	6	1.27343	1	Unmonitored	7	Respondents Nonrespondents	0.95% 0.26%	72.2% 33.5%	25.6% 5.7%	5.3% 0.0%	24.7% 4.6%	44.3% 89.8%
	18	1.36303	2	Unmonitored	4	Respondents Nonrespondents	5.82% 2.11%	54.5% 64.1%	14.9% 11.5%	0.9% 0.0%	19.6% 9.4%	64.6% 79.1%
Louisiana	23	1.29525	1	Unmonitored	7	Respondents Nonrespondents	0.72% 0.21%	56.8% 57.2%	25.3% 0.0%	71.1% 100%	3.6% 0.0%	0.0% 0.0%
Maryland	2	1.32123	1	Both	1	Respondents Nonrespondents	3.04% 0.98%	53.6% 56.4%	4.2% 43.7%	79.4% 56.3%	12.1% 0.0%	4.3% 0.0%
Michigan	3	1.27354	1	Unmonitored	1	Respondents Nonrespondents	0.69% 0.19%	65.3% 59.2%	0.0% 0.0%	80.9% 100%	19.1% 0.0%	0.0% 0.0%
	4	1.33955	1	Monitored	1	Respondents Nonrespondents	1.01% 0.34%	58.9% 39.7%	0.0% 0.0%	97.5% 100%	2.5% 0.0%	0.0% 0.0%
Minnesota	5	1.27655	1	Unmonitored	3	Respondents Nonrespondents	1.04% 0.29%	54.82% 56.4%	92.1% 100%	0.0% 0.0%	3.5% 0.0%	4.4% 0.0%
New York	2	1.30597	1	Monitored	1,2	Respondents Nonrespondents	3.81% 1.16%	59.2% 67.7%	24.4% 5.4%	35.5% 39.2%	38.5% 55.4%	1.6% 0.0%

---

* Age class 1 consists of students born in September 1977 or earlier. All other students are in age class 2.

BEST COPY AVAILABLE

Table 7-10 (continued)
Grade 8 Student Nonresponse Adjustment Classes with Adjustment Factors Greater than 1.25

State	Class	Nonresponse Adjustment	Age Class*	Monitor Status	Types of Locale	Response Status	Percent of State Population	Percent Male	Percent Race/Ethnicity			
									White	Black	Hispanic	Other
Ohio	1	1.37678	1	Unmonitored	2	Respondents	0.80%	58.0%	79.3%	8.3%	4.2%	8.3%
						Nonrespondents	0.30%	66.6%	88.9%	11.1%	0.0%	0.0%
	6	1.27871	1	Monitored	1,2	Respondents	1.98%	62.0%	19.4%	66.2%	12.8%	1.5%
						Nonrespondents	0.55%	66.8%	43.2%	49.4%	0.0%	7.4%
	7	1.27871	1	Unmonitored	3,4	Respondents	0.81%	69.7%	91.0%	0.0%	3.8%	5.2%
						Nonrespondents	0.23%	48.2%	100%	0.0%	0.0%	0.0%
Oklahoma	2	1.48880	1	Monitored	1,2	Respondents	1.34%	72.9%	84.4%	9.3%	3.0%	3.2%
						Nonrespondents	0.66%	24.5%	95.1%	4.9%	0.0%	0.0%
	5	1.65882	1	Both	1,2	Respondents	1.10%	54.7%	34.4%	39.4%	11.7%	14.5%
						Nonrespondents	0.72%	69.5%	41.9%	28.9%	6.9%	22.3%
	6	1.44623	1	Both	3	Respondents	0.80%	45.6%	61.6%	30.9%	7.4%	0.0%
						Nonrespondents	0.36%	70.0%	33.2%	60.0%	0.0%	6.9%
	8	1.43160	1	Both	3,4	Respondents	0.94%	68.2%	83.1%	8.8%	0.0%	8.2%
						Nonrespondents	0.41%	70.8%	89.9%	10.1%	0.0%	0.0%
	9	1.25685	1	Unmonitored	6	Respondents	1.64%	59.4%	60.0%	8.8%	3.5%	27.7%
						Nonrespondents	0.42%	65.4%	28.7%	8.4%	0.0%	62.9%
	11	1.31678	1	Unmonitored	6	Respondents	1.58%	54.7%	62.4%	11.7%	8.7%	17.3%
						Nonrespondents	0.50%	65.8%	56.2%	6.8%	6.8%	30.1%
	13	1.31857	1	Unmonitored	5,6	Respondents	2.11%	63.6%	85.0%	1.6%	9.7%	3.7%
						Nonrespondents	0.67%	69.4%	84.8%	0.0%	0.0%	15.2%
	14	1.33986	1	Monitored	6	Respondents	1.49%	55.2%	70.5%	5.5%	16.2%	7.9%
						Nonrespondents	0.51%	54.0%	92.3%	0.0%	0.0%	7.7%
	15	1.64300	1	Both	7	Respondents	1.25%	53.1%	71.6%	0.0%	9.2%	19.2%
						Nonrespondents	0.80%	75.2%	66.2%	0.0%	0.0%	33.8%

---

* Age class 1 consists of students born in September 1977 or earlier. All other students are in age class 2.

Table 7-10 (continued)
Grade 8 Student Nonresponse Adjustment Classes with Adjustment Factors Greater than 1.25

State	Class	Nonresponse Adjustment	Age Class*	Monitor Status	Types of Locale	Response Status	Percent of State Population	Percent Male	Percent Race/Ethnicity			
									White	Black	Hispanic	Other
Oklahoma	16	1.31762	1	Both	7	Respondents	2.07%	65.8%	74.8%	2.1%	10.7%	12.3%
						Nonrespondents	0.66%	64.4%	85.6%	0.0%	0.0%	14.4%
	17	1.38314	1	Both	7	Respondents	1.94%	68.0%	85.0%	0.0%	1.4%	13.7%
						Nonrespondents	0.74%	43.9%	89.4%	0.0%	0.0%	10.6%
	21	1.29275	2	Monitored	1,2	Respondents	1.81%	44.7%	68.9%	13.8%	10.8%	6.6%
						Nonrespondents	0.53%	52.5%	67.6%	13.5%	5.3%	13.5%
	23	1.43279	2	Monitored	1,2	Respondents	1.54%	34.1%	36.3%	46.7%	7.3%	9.7%
						Nonrespondents	0.66%	53.0%	53.2%	30.6%	4.1%	12.0%
	28	1.39573	2	Both	3,4	Respondents	2.41%	47.6%	88.3%	3.4%	3.2%	5.1%
						Nonrespondents	0.95%	47.4%	87.1%	8.6%	0.0%	4.3%
	33	1.30994	2	Unmonitored	5,6	Respondents	3.23%	50.8%	91.4%	0.0%	2.1%	6.5%
						Nonrespondents	1.00%	45.0%	89.0%	3.4%	0.0%	7.6%
	34	1.30706	2	Monitored	6	Respondents	2.81%	36.5%	78.8%	4.3%	5.5%	11.4%
						Nonrespondents	0.86%	36.9%	54.7%	9.1%	4.5%	31.7%
	40	1.26144	2	Monitored	7	Respondents	1.96%	50.8%	96.2%	0.0%	1.8%	2.0%
						Nonrespondents	0.51%	73.3%	96.1%	0.0%	0.0%	3.9%
Pennsylvania	1	1.28832	1	Both	2	Respondents	0.95%	49.3%	76.8%	11.5%	3.5%	8.2%
						Nonrespondents	0.27%	23.9%	88.2%	11.8%	0.0%	0.0%
Rhode Island	6	1.26521	1	Unmonitored	3	Respondents	1.49%	72.8%	73.1%	1.8%	21.7%	3.4%
						Nonrespondents	0.39%	64.8%	74.5%	12.7%	0.0%	12.7%
South Carolina	13	1.28036	1	Unmonitored	6	Respondents	0.98%	71.0%	27.7%	57.1%	7.6%	7.6%
						Nonrespondents	0.27%	54.5%	32.7%	67.3%	0.0%	0.0%
Tennessee	3	1.25992	1	Unmonitored	1	Respondents	1.14%	45.2%	65.4%	31.7%	2.9%	0.0%
						Nonrespondents	0.30%	35.3%	72.9%	27.1%	0.0%	0.0%

---

* Age class 1 consists of students born in September 1977 or earlier. All other students are in age class 2.

Table 7-10 (continued)
Grade 8 Student Nonresponse Adjustment Classes with Adjustment Factors Greater than 1.25

State	Class	Nonresponse Adjustment	Age Class*	Monitor Status	Types of Locale	Response Status	Percent of State Population	Percent Male	Percent Race/Ethnicity			
									White	Black	Hispanic	Other
Virginia	15	1.28708	1	Both	6	Respondents	1.05%	52.6%	39.3%	51.5%	3.7%	5.5%
						Nonrespondents	0.30%	68.0%	48.7%	51.3%	0.0%	0.0%
Wisconsin	15	1.26195	1	Monitored	7	Respondents	1.24%	68.8%	88.2%	0.0%	0.0%	11.8%
						Nonrespondents	0.32%	87.6%	85%	0.0%	0.0%	15.0%
West Virginia	9	1.41377	1	Monitored	5,6	Respondents	0.73%	59.8%	92.7%	7.3%	0.0%	0.0%
						Nonrespondents	0.30%	82.4%	100%	0.0%	0.0%	0.0%
	14	1.26926	1	Unmonitored	7	Respondents	2.51%	61.5%	94.4%	1.0%	3.6%	1.0%
						Nonrespondents	0.68%	88.6%	100%	0.0%	0.0%	0.0%

---

* Age class 1 consists of students born in September 1977 or earlier. All other students are in age class 2.

- *The percentage of students who are male, weighted by the base weight for each student adjusted for school nonresponse.* This estimates the proportion of students who are male in the subpopulation represented by the sample students.

- *The percentages of students who are White, Black, Hispanic, or of another race/ethnicity.* Again these percentages are weighted by the students' base weights, adjusted for school nonresponse.

The table shows that assessed and absent students have similar characteristics within nonresponse adjustment classes. A notable feature is that most of the cases involving adjustment factors in excess of 1.25 occur within classes in which the students are in age class 1—that is, relatively old for their grade. Since both the respondents and nonrespondents share this characteristic, this is not in itself a source of nonresponse bias. The potential for bias arises because of the possibility that, within this group, the respondents differ from the nonrespondents.

Note that invariably within a cell the size of the population represented by the nonrespondents is relatively small. Thus it is not likely in any state that substantial nonresponse bias could be arising from the nonresponse within a single cell. Rather, if such bias is occurring, it must be aggregated across a number of cells having varying characteristics except perhaps for the fact that they involve students of above average age. The small number of nonrespondents within each cell (often as few as five or six) makes it difficult to compare the characteristics of nonrespondents with those of respondents and to characterize the nonrespondents' distributions of gender, race/ethnicity, and median household income.

Of particular note in these tables is the presence of a large number of cells for Oklahoma with adjustment factors in excess of 1.25. This occurs because Oklahoma is the only state that required written parental consent before a selected student could participate in the assessment. This requirement resulted in much greater student nonresponse overall than in other states. What the results in Tables 7-9 and 7-10 suggest is that this nonresponse is very widely distributed across the various adjustment classes, and is not concentrated among particular types of students. This lessens (but does not eliminate) the likelihood that the high levels of student nonresponse in Oklahoma have introduced substantial nonresponse bias.

7.5 Variation in Weights

After completion of the weighting steps, an analysis was conducted of the distribution of the final student weights in each state. The analysis was intended to check that the various weight components had been derived properly in each state and to examine the impact of the variability of the sample weights on the precision of the sample estimates, both for the state as a whole and for major subgroups within the state.

The analysis was conducted by looking at the distribution of the final student weights, both for the approximately 2,500 assessed students in each state, grade, and subject, and for subgroups defined by age, gender, race/ethnicity, level of urbanicity, and level of parents' education. Two key aspects of the distribution were considered in each case: the coefficient of variation (equivalently, the relative variance) of the weight distribution; and the presence of outliers (i.e., cases whose weights were several standard deviations away from the median weight).

It was important to examine the coefficient of variation of the weights because a large coefficient of variation reduces the effective size of the sample. Assuming that the variables of interest for individual students are uncorrelated with the weights of the students, the sampling variance of an estimated average or aggregate is approximately $\{1+\left(\frac{C}{100}\right)^2\}$ times as great as the corresponding sampling variance based on a self-weighting sample of the same size, where C is the coefficient of variation of the weights expressed as a percent. Outliers, or cases with extreme weights, were examined because the presence of such an outlier was an indication of the possibility that an error was made in the weighting procedure, and because it was likely that a few extreme cases would contribute substantially to the size of the coefficient of variation.

In most states, the coefficients of variation were 35 percent or less, both for the whole sample and for all major subgroups. This means that the quantity $\{1+\left(\frac{C}{100}\right)^2\}$ was generally below 1.1, and the variation in sampling weights had little impact on the precision of sample estimates.

Large student weights were observed in a few states. These extreme weights generally affected those students in schools for which the grade enrollment available at the time of sample selection proved to be several-fold short of the actual enrollment. An evaluation was made of the impact of trimming these largest weights back to a level consistent with the remaining large weights found in the state. Such a procedure produced some reduction in the size of the coefficient of variation. It was sufficiently modest in each case, however, that we judged that the potential for the introduction of bias through trimming, when combined with the considerable effort required to implement an appropriate trimming procedure, was such that it was preferable not to apply any trimming to the weights in these states. The analyses conducted confirmed that weight components had been calculated and combined correctly, and it was concluded that weight trimming should not be undertaken. Note, however, that weight trimming of school base weights had already been applied in a few cases, prior to the analyses discussed here (see section 7.2.2).

7.6 Calculation of Replicate Weights

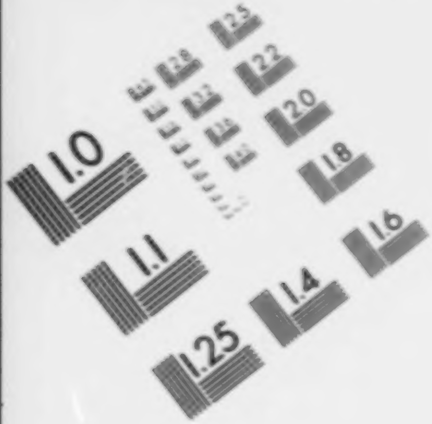
A method known as jackknife replication was used to estimate the sampling variance of statistics derived from the full sample. The process of replication involves repeatedly selecting portions of the sample to calculate the statistic of interest; the resultant estimates are known as replicate estimates. The variability among the calculated replicate estimates is then used to obtain the sampling variance of the full-sample estimate. The process of forming the replicate estimates is described below.

AIIM

Association for Information and Image Management

1100 Wayne Avenue, Suite 1100
Silver Spring, Maryland 20910

301/587-8202

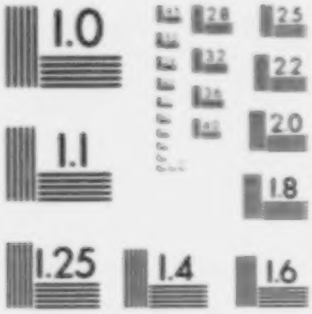

1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0

1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5

Centimeter

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
mm
1
2
3
4
5

Inches


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0

1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5

1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5

MANUFACTURED TO AIIM STANDARDS
BY APPLIED IMAGE, INC.

7.6.1 Defining Replicate Groups for Variance Estimation

To form replicates for variance estimation, the sampled cluster/hits in each Cluster Type 2 or 3 state (that is, those states where not all schools were selected) were sorted by monitor status, new-school status within monitor status, and finally by selection order within new-school status. The selection order used to form the replicate groups reflected the implicit stratification used in the selection of the sample of schools (see section 3.4.4). Within the sorted file, the basic algorithm for forming the replicate groups was to pair successive cluster/hits, separately within the two monitor status categories. A monitored cluster/hit was always paired with a monitored cluster/hit, and an unmonitored cluster/hit was always paired with an unmonitored cluster/hit. All members (schools) of a cluster/hit received the same pair code, and a substitute school received the pair code of the school it replaced. Double-session substitute schools were in effect assigned two pair codes, one corresponding to the original participating school and the other corresponding to the refusing school for which the extra sessions were conducted.

Since the schools in the Cluster Type 1 states were certainty schools, they were sorted and paired differently. First, each school was assigned a "half-group" code corresponding to the expected number of students selected from the school. For the fourth-grade sample (not including Guam and the Virgin Islands), the value of the half-group code was set to 1 if the expected number of sample students in the school was less than 90; otherwise, the value of the half-group code was set to 2. For the eighth-grade sample (not including Guam and the Virgin Islands), the value of the half-group code was set to 1 if the expected number of sample students in the school was less than 45; 2 if the expected number of sample students in the school was between 45 and 74, inclusive; and 4 if the expected number of sample students in the school was 75 or greater. For schools in Guam, the values of the half-group code ranged from 2 to 8, depending on the estimated grade enrollment of the school; for schools in the Virgin Islands, the values of the half-group code ranged from 2 to 16, depending on the estimated grade enrollment of the school. After assignment of the half-group codes, the schools within each Cluster Type 1 state were sorted by monitor status, half-group code (descending order) within monitor status, and by the estimated grade enrollment of the school within half-group code. Note that the half-group code essentially specifies the number of variance estimation units to be created from the school. For example, two clusters of students (i.e., variance-estimation units) were created from each school having a half-group code of 2, four clusters of students (i.e., variance-estimation units) were created from each school having a half-group code of 4; and so on. Each variance-estimation unit was a systematic sample of students within the school, and successive variance-estimation units in the sorted file were paired to define the replicates.

In some instances, there was an odd number of cluster/hits (in the case of Cluster Type 2 or 3 states) or variance-estimation units (in the case of Cluster Type 1 states) within a monitor-status category. If this occurred, the last "pair" within the monitor-status category actually consisted of three cluster/hits or variance-estimation units. In general, a single replicate was defined by randomly dropping a member (i.e., either a cluster/hit or variance-estimation unit) of a given pair and then reweighting the remaining sample elements to compensate for the dropped unit. If the pair consisted of three units, two groups of two units each were randomly retained to form two replicates.

The number of replicates formed in this manner depended on the number of pairs formed. Based on statistical and computer processing requirements, it was decided that 56 replicates would be sufficient for the variance calculations. In a few states, there were more than 56 initial pairs using the procedures described above. In these states, it was necessary to combine some of the initial replicate groups to reduce the total number of replicates. In general, the goal was to combine an initial pair with another pair consisting of dissimilar schools within the same monitor-status category.

In some states, fewer than 56 replicates were formed. In order to provide a uniform total of 56 replicates, additional sets of replicate weights were created simply by setting the additional sets equal to the set of full-sample weights. This procedure is unbiased and produces appropriate jackknifed sampling errors, while giving uniformity across states in the number of replicate weights.

7.6.2 School-level Replicate Weights

As mentioned above, each replicate sample had to be reweighted to compensate for the dropped unit(s) defining the replicate. For the Cluster Type 2 and 3 states, this reweighting was done in two stages. At the first stage, the *i*th school/hit included in a particular replicate *r* was assigned a replicate-specific school/hit base weight defined as follows:

$$W^{sch}_{(r)i} = K_r W^{sch}_i$$

where W^{sch}_i is the full-sample base weight for school/hit i, and

$$K_r = \begin{cases} 1.5 & \text{if school/hit } i \text{ was contained in a "pair" consisting of 3 units from which} \\ & \text{the complementary member was dropped to form replicate } r, \\ 2 & \text{if school/hit } i \text{ was contained in a pair consisting of 2 units from which the} \\ & \text{complementary member was dropped to form replicate } r, \\ 0 & \text{if school/hit } i \text{ was dropped to form replicate } r, \\ 1 & \text{otherwise.} \end{cases}$$

Using the replicate-specific school/hit base weights, $W^{sch}_{(r)i}$, the school-level nonresponse weighting adjustments as described in section 7.3.3 were recalculated for each replicate *r*. That is, the school-level nonresponse adjustment factor for schools in replicate *r* and adjustment class *h* was computed as:

$$F^{(1)}_{(r)h} = \frac{\sum_{i \in C_h} W^{sch}_{(r)hi} E_{hi}}{\sum_{i \in C_h} W^{sch}_{(r)hi} E_{hi} \delta_{(r)hi}}$$

where

C_h = the subset of school/hit records in adjustment class *h*;

$W^{sch}_{(r)hi}$ = the replicate-*r* base weight of the *i*th school/hit in class *h*;

E_{hi} = the QED grade enrollment for the *i*th school/hit in class *h*;

$$\delta_{(r)hi} = \begin{cases} 1 & \text{if the } i\text{th school/hit in replicate } r \text{ and adjustment class } h \text{ participated in the assessments; and} \\ 0 & \text{otherwise.} \end{cases}$$

The replicate-specific nonresponse-adjusted school/hit weight for the *i*th school in class *h* in replicate *r* was then computed as:

$$W^{adj}_{(r)hi} = F^{(1)}_{(r)h} W^{sch}_{(r)hi} \ .$$

7.6.3 Student-level Replicate Weights

For the Cluster Type 2 and 3 states, replicate-specific adjusted student base weights were calculated by multiplying the replicate-specific adjusted school/hit weights as described above by the corresponding within-school student weights. That is, following the procedures in section 7.3.5, the adjusted student base weight for the *j*th student in adjustment class *k* in replicate *r* was initially computed as:

$$W^{(2)}_{(r)kij} = W^{adj}_{(r)hi} W^{within}_{ij}$$

where

$W^{(2)}_{(r)hi}$ = the nonresponse-adjusted school/hit weight for school/hit *i* in school adjustment class *h* and replicate *r*;

W^{within}_{ij} = the within-school weight for the *j*th student in school *i*.

For the Cluster Type 1 states, the school-level nonresponse adjustment was not replicated since the schools in such states were selected with certainty. In this case, the replicate-specific adjusted student base weight for the *j*th student in adjustment class *k* in replicate *r* was calculated as:

$$W^{(2)}_{(r)kij} = W^{adj}_{hi} W^{within}_{(r)ij}$$

where

W^{adj}_{hi} = the overall nonresponse-adjusted school/hit weight for school/hit *i* in school adjustment class *h*;

$W^{within}_{(r)ij}$ = the replicate-specific within-school weight for the *j*th student in school *i*

= $K_r W^{within}_{ij}$

The factor K_r in the above expression for the replicate-specific within-school weight compensates for the units dropped out in any given replicate (see section 7.6.1) and is defined by:

$$K_r = \begin{cases} 2 & \text{for the students in school } i \text{ who were in a pair from which the complementary variance-estimation unit was dropped to form replicate } r, \\ 0 & \text{for the students in the variance-estimation unit that was dropped to form replicate } r, \\ 1 & \text{otherwise.} \end{cases}$$

The final replicate-specific student weights were then obtained by applying the student nonresponse adjustment procedures (see section 7.3.5) to each set of replicate student weights. Let $F^{(2)}_{(r)k}$ denote the student-level nonresponse adjustment factor for replicate *r* and adjustment class *k*. For the Cluster Type 2 and 3 states, the final replicate-*r* student weight for student *j* in school *i* in adjustment class *k* was calculated as:

$$W^{final}_{(r)kij} = F^{(2)}_{(r)k} W^{sch}_{(r)ki} W^{within}_{ij} \quad .$$

For the Cluster Type 1 states, the corresponding final replicate-*r* student weight for student *j* in school *i* in adjustment class *k* was calculated as:

$$W^{final}_{(r)kij} = F^{(2)}_{(r)k} W^{sch}_{ki} W^{within}_{ij} \quad .$$

Estimates of the variance of sample-based estimates were calculated as follows:

Let $\hat{x} = \sum_{i \in j}^{n} W^{final}_{kij} x_{kij}$ denote an estimated total based on the full sample, and let $\hat{x}_{(r)}$ denote the corresponding estimate based on replicate *r*. The jackknife variance estimate of $\hat{x}$ was calculated as:

$$var_{JK}(\hat{x}) = \sum_{r=1}^{R} (\hat{x}_{(r)} - \hat{x})^2 \, ,$$

where *R* is the number of replicates.

7.7 Calculation of School Weights

Since schools in the Cluster Type 1 states were selected with certainty, the "school/hit" weights described in section 7.3.3 can be used to estimate school-level characteristics and aggregates. However, these school/hit weights are not appropriate for the Cluster Type 2 and 3 states because large schools had a chance of being selected more than once in the sampling process. To compensate for the possibility of multiple selections, schools in the Cluster Type 2 and 3 states were assigned school weights, W_{hs}^{adj}, equal to:

$$W_{hs}^{adj} = \sum_{i \in s} W_{hi}^{adj}$$

where W_{hi}^{adj} is the adjusted school/hit weight for school/hit *i* in adjustment class *h*, and where the sum extends over the school/hits corresponding to school *s*. Similarly, the replicate-specific school weights were computed as:

$$W_{(r)hs}^{adj} = \sum_{i \in s} W_{(r)hi}^{adj}$$

where $W_{(r)hi}^{adj}$ is the replicate-specific adjusted school/hit weight defined in section 7.6.2.

Chapter 8

THEORETICAL BACKGROUND AND PHILOSOPHY OF NAEP SCALING PROCEDURES

Eugene G. Johnson, Robert J. Mislevy, and Neal Thomas

Educational Testing Service

8.1 OVERVIEW

The primary method by which results from the Trial State Assessment are disseminated is scale-score reporting. With scaling methods, the performance of a sample of students in a subject area or subarea can be summarized on a single scale or series of subscales even when different students have been administered different items. This chapter presents an overview of the scaling methodologies employed in the analyses of the data from NAEP surveys in general and from the Trial State Assessment of mathematics in particular. Details of the scaling procedures specific to the Trial State Assessment are presented in Chapter 9.

8.2 BACKGROUND

The basic information from an assessment consists of the responses of students to the items presented in the assessment. For NAEP, these items are generated to measure performance on sets of objectives developed by nationally representative panels of learning area specialists, educators, and concerned citizens. Satisfying the objectives of the assessment and ensuring that the tasks selected to measure each goal cover a range of difficulty levels typically requires a large number of items. The Trial State Assessment of mathematics required 175 items at grade 4 and 205 items at grade 8. To reduce student burden, each assessed student was presented only a fraction of the full pool of items using multiple matrix sampling procedures.

The most direct manner of presenting the assessment results is to report percent correct statistics for each item. However, because of the vast amount of information, separate results for each of the items in the assessment pool hinders the comparison of the general performance of subgroups of the population. Item-by-item reporting ignores overarching similarities in trends and subgroup comparisons that are common across items.

It is useful to view the assessed items as random representatives of a conceptually infinite pool of items within the same domain and of the same type. In this random item concept, a set of items is taken to represent the domain of interest. An obvious measure of achievement within a domain of interest is the average percent correct across all presented items within that domain. The advantage of averaging is that it tends to cancel out the effects of peculiarities in items that can affect item difficulty in unpredictable ways. Furthermore, averaging makes it possible to compare more easily the general performances of subpopulations.

Despite their advantages, there are a number of significant problems with average percent correct scores. First, the interpretation of these results depends on the selection of the items; the selection of easy or difficult items could make student performance appear to be overly high or low. Second, the average percent correct metric is related to the particular items comprising the average, so that direct comparisons in performance between subpopulations require that those subpopulations have been administered the same set of items. Third, because this approach limits comparisons to percents correct on specific sets of items, it provides no simple way to report trends over time when the item pool changes. Finally, direct estimates of statistics such as the proportion of students who would respond correctly to 80 percent of the items in the pool are not possible when every student is administered only a fraction of the item pool. While the mean percent correct across all items in the pool can be readily obtained (as the average of the individual item percent correct statistics), distributional statistics, such quantiles of the distribution of scores across the full set of items, cannot be readily obtained without additional assumptions.

These limitations can be overcome by the use of response scaling methods. If several items require similar skills, the regularities observed in response patterns can often be exploited to characterize both respondents and items in terms of a relatively small number of variables. These variables include a respondent-specific variable, called proficiency, which quantifies a respondent tendency to answer items correctly, and item-specific variables, which indicate characteristics of the item such as its difficulty, ability to distinguish between individuals with different levels of proficiency, and the chances of a very low proficiency respondent correctly answering the item. (These variables are discussed in more detail in the next section). When combined through appropriate mathematical formulas, these variables capture the dominant features of the data. Furthermore, all students can be placed on a common scale, even though none of the respondents take all of the items within the pool. Using the scale, it becomes possible to discuss distributions of proficiency in a population or subpopulation and to estimate the relationships between proficiency and background variables.

It is important to point out that any procedure of aggregation, from a simple average to a complex multidimensional scaling model, highlights certain patterns at the expense of other potentially interesting patterns that may reside within the data. Every item in a NAEP survey is of interest and can provide useful information about what young Americans know and can do. The choice of an aggregation procedure must be driven by a conception of just which patterns are salient for a particular purpose.

The scaling for the Trial State Assessment was carried out separately within the five mathematics content areas specified in the framework and for items designed to measure skills in estimation. This scaling within subareas was done because it was anticipated that different patterns of performance might exist for these essential subdivisions of the subject area. Each content area scale corresponded to one of five content areas: Numbers and Operations; Measurement; Geometry; Data Analysis, Statistics, and Probability; and Algebra and Functions. By creating a separate scale for each of these content areas, potential differences in subpopulation performance between the content areas are maintained. The separate estimation scale was created from an additional, special set of items measuring estimation skills (see section 2.6). The separate estimation scale allows for the measurement of potential performance differences within that skill area relative to performance on the other scales. Analyses of the results for the separate scales from the 1992 Trial State Assessment and national mathematics assessment have shown that the separate scales provide additional information that a single scale cannot—for example, gender differences in mathematics performance by type of scale.

The creation of a series of separate scales to describe mathematics performance does not preclude the reporting of an overall mathematics composite as a single index of overall mathematics performance. A composite is computed as the weighted average of the five content area scales, where the weights correspond to the relative importance given to each content area as defined by the objectives. The composite provides a global measure of performance within the subject area, while the constituent content area scales allow the measurement of important interactions within educationally relevant subdivisions of the subject area.

8.3 SCALING METHODOLOGY

This section reviews the scaling models employed in the analyses of data from the Trial State Assessment of mathematics and the 1992 national mathematics assessment, and the "plausible values" methodology that allows such models to be used with NAEP's sparse item-sampling design. The reader is referred to Mislevy (1991) for an introduction to plausible values methods and a comparison with standard psychometric analyses, to Mislevy, Johnson and Muraki (1992) and Beaton and Johnson (1992) for additional information on how the models are used in NAEP, and to Rubin (1987) for the theoretical underpinnings of the approach.

While the NAEP procedures were developed explicitly to handle the characteristics of NAEP data, they build on other research, and are paralleled by other researchers. See, for example: Dempster, Laird, and Rubin (1977); Little and Rubin (1983,1987); Andersen (1980); Engelen (1987); Hoijtink (1991); Laird (1978); Lindsey, Clogg, and Grego (1991); Zwinderman (1991); Tanner and Wong (1987); and Rubin (1991).

The 175 mathematics items administered at grade 4 and the 205 items administered at grade 8 in the Trial State Assessment were also administered to students of the same grades in the national mathematics assessment. However, because the administration procedures differed, the Trial State Assessment data was scaled independently from the national data. The national data also included results for students in grade 12. Details of the scaling of the Trial State Assessment and the subsequent linking to the results from the national mathematics assessment are provided in Chapter 9.

8.3.1 The Scaling Models

Three distinct scaling models were used in the analysis of the data from the Trial State Assessment. Each of the models are based on item response theory (IRT; e.g., Lord, 1980). Each is a "latent variable" model, defined separately for each of the scales, and quantifying respondents' tendencies to provide correct answers to the items contributing to a scale as a function of a parameter that is not directly observed, called proficiency on the scale.

A three-parameter logistic (3PL) model was used for the multiple-choice items. The fundamental equation of the 3PL model is the probability that a person whose proficiency on scale k is characterized by the *unobservable* variable θ_k will respond correctly to item j:

$$P(X_j = 1|\theta_k, a_j, b_j, c_j) = C_j + \frac{(1 - C_j)}{1 + \exp[-1.7a_j\,(\theta_k - b_j)]} \tag{8.1}$$

$$\equiv P_{j1}(\theta_k) \quad ,$$

where

x_j	is the response to item j, 1 if correct and 0 if not;
a_j	where $a_j > 0$, is the slope parameter of item j, characterizing its sensitivity to proficiency;
b_j	is the threshold parameter of item j, characterizing its difficulty; and
c_j	where $0 \leq c_j < 1$, is the lower asymptote parameter of item j, reflecting the chances of students of very low proficiency selecting the correct option.

Further define the probability of an incorrect response to the item as

$$P_{j0} = P(x_j = 0|\theta_k, a_j, b_j, c_j) = 1 - P_{j1}(\theta_k) \tag{8.2}$$

A two-parameter logistic (2PL) model was used for short constructed-response items, which were scored correct or incorrect. The form of the 2PL model is the same as equations (8.1) and (8.2) with the c_j parameter fixed at zero.

In addition to the multiple-choice and short constructed-response items, a number of extended constructed-response items (5 at grade 4 and 6 at grade 8) were presented in the Trial State and national assessment. Each of these items was scored on a multipoint scale with potential scores ranging from 0 to 4. Additionally, as discussed in Chapter 9, certain sets of items consisting of highly correlated parts were combined into "testlets" (Wainer & Kiely, 1987) where the score assigned to a testlet was the number of constituent parts answered correctly. Items which are scored on a multipoint scale are referred to as polytomous items, in contrast with the multiple-choice and short constructed-response items, which are scored correct/incorrect and referred to as dichotomous items.

The polytomous items were scaled using a generalized partial credit model (Muraki, 1992). The fundamental equation of this model is the probability that a person with proficiency θ_k on scale k will have, for the jth polytomous item, a response x_j that is scored in the ith of m_j ordered score categories:

$$P(X_j = i|\theta_k, a_j, b_j, d_{j1}, \ldots, d_{jm_j-1}) = \frac{\exp(\sum_{v=0}^{i} 1.7a_j(\theta_k - b_j + d_{jv}))}{\sum_{g=0}^{m_j-1} \exp(\sum_{v=0}^{g} 1.7a_j(\theta_k - b_j + d_{jv}))} \tag{8.3}$$

$$= P_{ji}(\theta_k)$$

where

m_j is the number of categories in the response to item *j*

x_j is the response to item *j*, with possibilities $0,1,\ldots,m_j-1$

a_j is the slope parameter

b_j is the item location parameter characterizing overall difficulty

d_{ji} is the category *i* threshold parameter (see below).

Indeterminacies in the parameters of the above model are resolved by setting $d_{j0} = 0$ and setting $\sum_{i=1}^{m_j-1} d_{ji} = 0$. Muraki (1992) points out that $b_j - d_{ji}$ is the point on the θ_k scale at which the plots of $P_{j,i-1}(\theta_k)$ and $P_{ji}(\theta_k)$ intersect and so characterizes the point on the θ_k scale at which the response to item *j* has the highest probability of incurring a change from response category *i*-1 to *i*.

When $m_j = 2$, so that there are two score categories (0,1), it can be shown that $P_{ji}(\theta_k)$ of equation 8.3 for i=0,1 corresponds respectively to $P_{j0}(\theta_k)$ and $P_{j1}(\theta_k)$ of the 2PL model (equations 8.1 and 8.2 with c_j=0).

A typical assumption of item response theory is the conditional independence of the probabilities of correct response by an individual to a set of items, given the individual's proficiency. That is, conditional on the individual's θ_k, the joint probability of a particular response pattern $\underline{x} = (x_1,\ldots,x_n)$ across a set of *n* items is simply the product of terms based on (8.1), (8.2), and (8.3):

$$P(\underline{x}|\theta_k, \textit{item parameters}) = \prod_{j=1}^{n} \prod_{i=0}^{m_j-1} P_{ji}(\theta_k)^{u_{ji}} \tag{8.4}$$

where $P_{ji}(\theta_k)$ is of the form appropriate to the type of item (dichotomous or polytomous), m_j is taken equal to 2 for the dichotomously scored items, and u_{ji} is an indicator variable defined by

$$u_{ji} = \begin{cases} 1 \text{ if response } x_j \text{ was in category } i \\ 0 \text{ otherwise.} \end{cases}$$

It is also typically assumed that response probabilities are conditionally independent of background variables ($\underline{y}$), given θ_k, or

$$P(\underline{x}|\theta_k, \mathit{item\ parameters}, \underline{y}) = p(\underline{x}|\theta_k, \mathit{item\ parameters}) \qquad (8.5)$$

After $\underline{x}$ has been observed, equation 8.4 can be viewed as a likelihood function, and provides a basis for inference about θ_k or about item parameters. Estimates of item parameters were obtained by the NAEP BILOG/PARSCALE program, which combines Mislevy and Bock's (1982) BILOG and Muraki and Bock's (1991) PARSCALE computer programs, and which concurrently estimates parameters for all items (dichotomous and polytomous). The item parameters are then treated as known in subsequent calculations. The parameters of the items constituting each of the separate scales were estimated independently of the parameters of the other scales. Once items have been calibrated in this manner, a likelihood function for the scale proficiency θ_k is induced by a vector of responses to any subset of calibrated items, thus allowing θ_k-based inferences from matrix samples.

Item parameter estimation was performed separately for the grade 4 and the grade 8 data. As stated previously, item parameter estimation was performed independently for the Trial State Assessment and for the national mathematics assessment. In both cases, the identical scale definitions were used.

In all NAEP IRT analyses, missing responses at the end of each block a student was administered were considered "not-reached," and treated as if they had not been presented to the respondent. Missing responses to dichotomous items before the last observed response in a block were considered intentional omissions, and treated as fractionally correct at the value of the reciprocal of the number of response alternatives. These conventions are discussed by Mislevy and Wu (1988). With regard to the handling of not-reached items, Mislevy and Wu found that ignoring not-reached items introduces slight biases into item parameter estimation to the degree that not-reached items are present and speed is correlated with ability. With regard to omissions, they found that the method described above provides consistent limited-information likelihood estimates of item and ability parameters under the assumption that respondents omit only if they can do no better than responding randomly.

Because the extended constructed-response items were always the last item in a block and because considerably more effort was required of the student to answer these items, nonresponse to an extended constructed-response item was considered an intentional omission (and scored as the lowest category, 0) unless the student also did not respond to the item immediately preceding that item. In that case, the extended constructed-response item was considered not reached and treated as if it had not been presented to the student.

Although the IRT models are employed in NAEP only to summarize performance, a number of checks are made to detect serious violations of the assumptions underlying the models (such as conditional independence), and, when warranted, remedial efforts are made to mitigate the effects of such violations on inferences. These checks include comparisons of empirical and theoretical item response functions to identify items for which the IRT model may provide a poor fit to the data.

Scaling areas in NAEP are determined *a priori* by considerations of content as collections of items for which overall performance is deemed to be of interest, as defined by the frameworks developed by the National Assessment Governing Board. A proficiency scale θ_k is defined *a priori* by the collection of items representing that scale. What is important, therefore, is that the models capture salient variation in the response data to effectively summarize the overall performance on the content area of the populations and subpopulations being assessed. Because of the *a priori* definition of the latent proficiency variable, departure from conditional independence tends to cancel out over items and does not seriously affect the estimation of whole group and subpopulation distributions, except when substantial differential item functioning (DIF) is found simultaneously for many items. NAEP has routinely conducted DIF analyses to guard against potential biases in making subpopulation comparisons based on the proficiency distributions.

The local independence assumption embodied in equation 8.4 implies that item response probabilities depend only on θ and the specified item parameters, and not on the position of the item in the booklet, the content of items around an item of interest, or the test-administration timing conditions. However, these effects are certainly present in any application. The practical question is whether inferences based on the IRT probabilities obtained via 8.4 are robust with respect to the ideal assumptions underlying the IRT model. Our experience with the 1986 NAEP reading anomaly has shown that for measuring small changes over time, changes in item context and speededness conditions can lead to unacceptably large random error components. These can be avoided by presenting items used to measure change in identical test forms, with identical timings and administration conditions. Thus, we do *not* maintain that the item parameter estimates obtained in any particular booklet configuration are appropriate for other conceivable configurations. Rather, we assume that the parameter estimates are context-bound. (For this reason, we prefer common population equating to common item equating whenever equivalent random samples are available for linking.) This is the reason that the data from the Trial State Assessment were calibrated separately from the data from the national NAEP—since the administration procedures differed somewhat between the Trial State Assessment and the national NAEP, the values of the item parameters could be different. Furthermore, to allow for the possibility that item parameters could change over time, the 1992 grade 8 calibration was conducted separately from the data from the 1990 grade 8 Trial State Assessment. Chapter 9 provides details on the procedures used to link the results of the 1992 Trial State Assessment to those of the 1992 national assessment and, hence, to those of the 1990 Trial State and National Assessments.

8.3.2 An Overview of Plausible Values Methodology

Item response theory was developed in the context of measuring individual examinees' abilities. In that setting, each individual is administered enough items (often 100 or more) to permit precise estimation of his or her θ, as a maximum likelihood estimate $\hat{\theta}$, for example. Because the uncertainty associated with each θ is negligible, the distribution of θ, or the joint distribution of θ with other variables, can then be approximated using individuals' $\hat{\theta}$ values as if they were θ values.

This approach breaks down in the assessment setting when, in order to provide broader content coverage in limited testing time, each respondent is administered relatively few items in a scaling area. The problem is that the uncertainty associated with individual θs is too large to ignore, and the features of the $\hat{\theta}$ distribution can be seriously biased as estimates of the θ distribution. (The failure of this approach was verified in early analyses of the 1984 NAEP reading survey; see Wingersky, Kaplan, & Beaton, 1987.) "Plausible values" were developed as a way to estimate key population features consistently, and approximate others no worse than standard IRT procedures would. A detailed development of plausible values methodology is given in Mislevy (1991). Along with theoretical justifications, that paper presents comparisons with standard procedures, discussions of biases that arise in some secondary analyses, and numerical examples.

The following provides a brief overview of the plausible values approach, focusing on its implementation in the Trial State Assessment analyses.

Let $\underline{y}$ represent the responses of all sampled examinees to background and attitude questions, along with design variables such as school membership, and let $\underline{\theta}$ represent the subscale proficiency values. If $\underline{\theta}$ were known for all sampled examinees, it would be possible to compute a statistic $t(\underline{\theta},\underline{y})$—such as a subscale or composite subpopulation sample mean, a sample percentile point, or a sample regression coefficient—to estimate a corresponding population quantity T. A function $U(\underline{\theta},\underline{y})$—e.g., a jackknife estimate—would be used to gauge sampling uncertainty, as the variance of t around T in repeated samples from the population.

Because the scaling models are latent variable models, however, $\underline{\theta}$ values are not observed even for sampled students. To overcome this problem, we follow Rubin (1987) by considering $\underline{\theta}$ as "missing data" and approximate $t(\underline{\theta},\underline{y})$ by its expectation given $(\underline{x},\underline{y})$, the data that actually were observed, as follows:

$$\begin{aligned} t^*(\underline{x},\underline{y}) &= E[t(\underline{\theta},\underline{y})|\underline{x},\underline{y}] \\ &= \int t(\underline{\theta},\underline{y})\, p(\underline{\theta}|\underline{x},\underline{y})\, d\underline{\theta} \ . \qquad (8.6) \end{aligned}$$

It is possible to approximate t^* using random draws from the conditional distribution of the scale proficiencies given the item responses x_i, background variables y_i, and model parameters for sampled student i. These values are referred to as "imputations" in the sampling literature, and "plausible values" in NAEP. The value of $\underline{\theta}$ for any respondent that would enter into the computation of t is thus replaced by a randomly selected value from their conditional distribution. Rubin (1987) proposes that this process be carried out several times—"multiple imputations"—so that the uncertainty associated with imputation can be quantified. The average of the results of, for example, M estimates of t, each computed from a different set of plausible values, is a Monte Carlo approximation of (8.6); the variance among them, B, reflects uncertainty due to not observing θ, and must be added to the estimated expectation of $U(\underline{\theta},\underline{y})$, which reflects uncertainty due to testing only a sample of students from the population. Section 8.4 explains how plausible values are used in subsequent analyses.

It cannot be emphasized too strongly that **plausible values are *not* test scores for *individuals*** in the usual sense. Plausible values are offered only as intermediary computations for calculating integrals of the form of equation 8.6, in order to estimate *population* characteristics. When the underlying model is correctly specified, plausible values will provide consistent estimates of population characteristics, even though they are not generally unbiased estimates of the proficiencies of the individuals with whom they are associated. The key idea lies in a contrast between plausible values and the more familiar θ estimates of educational measurement that are in some sense optimal for each examinee (e.g., maximum likelihood estimates, which are consistent estimates of an examinee's θ, and Bayes estimates, which provide minimum mean-squared errors with respect to a reference population): *Point estimates that are optimal for individual examinees have distributions that can produce decidedly nonoptimal (specifically, inconsistent) estimates of population characteristics* (Little & Rubin, 1983). Plausible values, on the other hand, are constructed explicitly to provide consistent estimates of population effects.

8.3.3 Computing Plausible Values in IRT-based Scales

Plausible values for each respondent i are drawn from the conditional distribution $p(\underline{\theta}_i | x_i, y_i, \Gamma, \Sigma)$, where Γ and Σ are regression model parameters defined in this subsection. This subsection describes how, in IRT-based scales, these conditional distributions are characterized, and how the draws are taken. An application of Bayes' theorem with the IRT assumption of conditional independence produces

$$p(\underline{\theta}_i | x_i, y_i, \Gamma, \Sigma) \propto P(x_i | \underline{\theta}_i, y_i, \Gamma, \Sigma)\, p(\underline{\theta}_i | y_i, \Gamma, \Sigma) = P(x_i | \underline{\theta}_i)\, p(\underline{\theta}_i | y_i, \Gamma, \Sigma) \ , \tag{8.7}$$

where, for vector-valued $\underline{\theta}_i$, $P(x_i | \underline{\theta}_i)$ is the product over scales of the *independent likelihoods* induced by responses to items within each scale, and $p(\underline{\theta}_i | y_i, \Gamma, \Sigma)$ is the multivariate—and generally nonindependent—*joint density* of proficiencies for the scales, conditional on the observed value y_i of background responses, and the parameters Γ and Σ. The scales are determined by the item parameter estimates that constrain the population mean to zero and standard deviation to one. The item parameter estimates are fixed and regarded as population values in the computation described in this subsection.

In the analyses of the data from the Trial State Assessment and the data from the national mathematics assessment, a normal (Gaussian) form was assumed for $p(\underline{\theta}_i | y_i \Gamma, \Sigma)$, with a common variance, Σ, and with a mean given by a linear model with slope parameters, Γ, based on the first 98 to 154 principal components of 258 (grade 4) and 303 (grade 8) selected main-effects and two-way interactions of the complete vector of background variables. The included principal components will be referred to as the *conditioning variables*, and will be denoted y^c. (The complete set of original background variables used in the Trial State Assessment analyses are listed in Appendix C.) The following model was fit to the data within each state:

$$\theta = \Gamma' y^c + \varepsilon \ , \tag{8.8}$$

where ε is normally distributed with mean zero and variance Σ. The number of principal components of the conditioning variables used for each state was sufficient to account for 90 percent of the total variance of the full set of conditioning variables (after standardizing each variable). As in regression analysis, Γ is a matrix each of whose columns is the *effects* for one scale and Σ is the matrix *variance of residuals* between subscales. By fitting the model (8.8) separately within each state, interactions between each state and the conditioning variables are automatically included in the conditional joint density of scale proficiencies.

Maximum likelihood estimates of Γ and Σ, denoted by $\hat{\Gamma}$ and $\hat{\Sigma}$, are obtained from Sheehan's (1985) MGROUP computer program using the EM algorithm described in Mislevy (1985). The EM algorithm requires the computation of the mean, $\bar{\underline{\theta}}_i$, and variance, Σ_i^p, of the posterior distribution in (8.7). These moments are computed using higher order asymptotic corrections (Thomas, 1992).

After completion of the EM algorithm, the plausible values are drawn in a three-step process from the joint distribution of the values of Γ for all sampled respondents. First, a value of Γ is drawn from a normal approximation to $P(\Gamma,\Sigma|x_i,y_i)$ that fixes Σ at the value $\hat{\Sigma}$, (Thomas, 1992). Second, conditional on the generated value of Γ (and the fixed value of $\Sigma = \hat{\Sigma}$), the mean, $\bar{\underline{\theta}}_i$, and variance, Σ_i^p, of the posterior distribution in equation 8.7 (i.e., $p(\underline{\theta}_i|x_i,y_i,\Gamma,\Sigma)$) are computed using the same methods applied in the EM algorithm. In the third step, the $\underline{\theta}_i$ are drawn independently from a multivariate normal distribution with mean $\bar{\underline{\theta}}_i$ and variance Σ_i^p, approximating the distribution in (8.7). These three steps are repeated five times producing five imputations of $\bar{\underline{\theta}}_i$ for each sampled respondent.

8.4 ANALYSES

When survey variables are observed without error from every respondent, standard variance estimators quantify the uncertainty associated with sample statistics from the only source of the uncertainty, namely the sampling of respondents. Item percents correct for NAEP cognitive items meet this requirement, but scale-score proficiency values do not. The IRT models used in their construction posit an unobservable proficiency variable θ to summarize performance on the items in the subarea. The fact that θ values are not observed even for the respondents in the sample requires additional statistical analyses to draw inferences about θ distributions and to quantify the uncertainty associated with those inferences. As described above, Rubin's (1987) multiple imputations procedures were adapted to the context of latent variable models to produce the plausible values upon which many analyses of the data from the Trial State Assessment were based. This section describes how plausible values were employed in subsequent analyses to yield inferences about population and subpopulation distributions of proficiencies.

8.4.1 Computational Procedures

Even though one does not observe the θ value of respondent i, one does observe variables that are related to it: x_i, the respondent's answers to the cognitive items he or she was administered in the area of interest, and y_i, the respondent's answers to demographic and background variables. Suppose one wishes to draw inferences about a number $T(\underline{\theta},Y)$ that could be calculated explicitly if the θ and y values of each member of the population were known. Suppose further that if θ values were observable, we would be able to estimate T from a sample of N pairs of θ and y values by the statistic $t(\underline{\theta},\underline{y})$ [where $(\underline{\theta},\underline{y}) = (\theta_1,y_1,\ldots,\theta_N,y_N)$], and that we could estimate the variance in t around T due to sampling respondents by the function $U(\underline{\theta},\underline{y})$. Given that observations consist of (x_i,y_i) rather than $(\underline{\theta}_i,y_i)$, we can approximate t by its expected value conditional on $(\underline{x},\underline{y})$, or

$$t^*(\underline{x},\underline{y}) = E[t(\underline{\theta},\underline{y})|\underline{x},\underline{y}] = \int t(\underline{\theta},\underline{y})\, p(\underline{\theta}|\underline{x},\underline{y})\, d\underline{\theta}\,.$$

It is possible to approximate t^* with random draws from the conditional distributions $p(\underline{\theta}_i|x_i,y_i)$, which are obtained for all respondents by the method described in section 8.3.3. Let $\underline{\theta}_m$ be the mth such vector of "plausible values," consisting of a multidimensional value for the latent variable of each respondent. This vector is a plausible representation of what the true $\underline{\theta}$ vector might have been, had we been able to observe it.

The following steps describe how an estimate of a scalar statistic $t(\underline{\theta},\underline{y})$ and its sampling variance can be obtained from M (>1) such sets of plausible values. (Five sets of plausible values are used in NAEP analyses of the Trial State Assessment.)

1) Using each set of plausible values $\hat{\underline{\theta}}_m$ in turn, evaluate t as if the plausible values were true values of $\underline{\theta}$. Denote the results $\hat{t}_m$ for $m=1,\ldots,M$.

2) Using the jackknife variance estimator defined in Chapter 7, compute the estimated sampling variance of $\hat{t}_m$ denoting the result U_m.

3) The final estimate of t is

$$t^* = \sum_{m=1}^{M} \frac{\hat{t}_m}{M}.$$

4) Compute the average sampling variance over the M sets of plausible values, to approximate uncertainty due to sampling respondents:

$$U^* = \sum_{m=1}^{M} \frac{U_m}{M}.$$

5) Compute the variance among the M estimates $\hat{t}_m$ to approximate uncertainty due to not observing θ values from respondents:

$$B_M = \sum_{m=1}^{M} \frac{(\hat{t}_m - t^*)^2}{(M-1)}$$

6) The final estimate of the variance of t^* is the sum of two components:

$$V = U^* + (1 + M^{-1})\, B_M.$$

Note: Due to the excessive computation that would be required, NAEP analyses did not compute and average jackknife variances over all five sets of plausible values, but only on the first set. Thus, in NAEP reports, U^* is approximated by U_1.

8.4.2 Statistical Tests

Suppose that if θ values were observed for sampled students, the statistic $(t - T)/U^{1/2}$ would follow a t-distribution with d degrees of freedom. Then the incomplete-data statistic $(t^* - T)/V^{1/2}$ is approximately t-distributed, with degrees of freedom given by

$$v = \frac{1}{\dfrac{f_M^2}{M-1} + \dfrac{(1-f_M)^2}{d}}$$

where f_M is the proportion of total variance due to not observing θ values:

$$f_M = (1+M^{-1})\, B_M / V_M.$$

When B_M is small relative to U^*, the reference distribution for incomplete-data statistics differs little from the reference distribution for the corresponding complete-data statistics. This is the case with main NAEP reporting variables. If, in addition, d is large, the normal approximation can be used to flag "significant" results.

For k-dimensional t, such as the k coefficients in a multiple regression analysis, each U_m and U^* is a covariance matrix, and B_M is an average of squares and cross-products rather than simply an average of squares. In this case, the quantity $(T-t^*)\, V^{-1}\, (T-t^*)'$ is approximately F distributed, with degrees of freedom equal to k and ν, with ν defined as above but with a matrix generalization of f_M:

$$f_M = (1+M^{-1})\ \mathrm{Trace}\ (B_M V_M^{-1})/k\ .$$

By the same reasoning as used for the normal approximation for scalar t, a chi-square distribution on k degrees of freedom often suffices.

8.4.3 Biases in Secondary Analyses

Statistics t^* that involve proficiencies in a scaled content area and variables included in the conditioning variables y^c, are consistent estimates of the corresponding population values T. Statistics involving background variables y that were *not* conditioned on, or relationships among proficiencies from *different* content areas, are subject to asymptotic biases whose magnitudes depend on the type of statistic and the strength of the relationships of the nonconditioned background variables to the variables that were conditioned on and to the proficiency of interest. That is, the large sample expectations of certain sample statistics need not equal the true population parameters.

The *direction* of the bias is typically to underestimate the effect of nonconditioned variables. For details and derivations see Beaton and Johnson (1990), Mislevy (1991), and Mislevy and Sheehan (1987, section 10.3.5). For a given statistic t^* involving one content area and one or more nonconditioned background variables, the *magnitude* of the bias is related to the extent to which observed responses x account for the latent variable θ, and the degree to which the nonconditioned background variables are explained by conditioning background variables. The first factor—conceptually related to test reliability—acts consistently in that greater measurement precision reduces biases in *all* secondary analyses. The second factor acts to reduce biases in certain analyses but increase it in others. In particular,

- High shared variance between conditioned and nonconditioned background variables *mitigates* biases in analyses that involve only proficiency and nonconditioned variables, such as marginal means or regressions.
- High shared variance *exacerbates* biases in regression coefficients of conditional effects for nonconditioned variables, when nonconditioned and conditioned background variables are analyzed jointly as in multiple regression.

The large number of background variables that have been included in the conditioning vector for the Trial State Assessment allows a large number of secondary analyses to be carried out with little or no bias, and mitigates biases in analyses of the marginal distributions of θ in nonconditioned variables. Kaplan and Nelson's analysis of the 1988 NAEP reading data (some results of which are summarized in Mislevy, 1991), which had a similar design and fewer conditioning variables, indicate that the potential bias for nonconditioned variables in multiple regression analyses is below 10 percent, and biases in simple regression of such variables is below 5 percent. Additional research (summarized in Mislevy, 1990) indicates that most of the bias reduction obtainable from conditioning on a large number of variables can be captured by instead conditioning on the first several principal components of the matrix of all original conditioning variables. This procedure was adopted for the Trial State Assessment by replacing the conditioning effects by the first K principal components, where K was selected so that 90 percent of the total variance of the full set of conditioning variables (after standardization) was captured. Mislevy (1991) shows that this puts an upper bound of 10 percent on the average bias for all analyses involving the original conditioning variables.

8.5 SCALE ANCHORING AND ACHIEVEMENT LEVELS

Since its beginning, a goal of NAEP has been to inform the public about what students in American schools know and can do. While the NAEP scales provide information about the distributions of proficiency for the various subpopulations, they do not directly provide information about the meaning of various points on the scale. Traditionally, meaning has been attached to educational scales by norm-referencing—that is, by comparing students at a particular scale level to other students. In contrast, NAEP scale anchors and achievement levels describe selected points on the scale in terms of the types of skills that are or should be exhibited by students scoring at that level. Both the scale anchoring and the achievement level processes were applied to the 1992 national NAEP mathematics composite. However, since the Trial State Assessment scales were linked to the national scales, the interpretations of the selected levels also apply to the Trial State Assessment.

As applied to the 1992 mathematics data, scale anchoring began by identifying four anchoring levels on the mathematics composite: 200, 250, 300, 350. The next step was to identify items that a large majority (at least 65 percent) of students at a given anchor level could answer correctly but that most students (at least 50 percent) at the next lower level answered incorrectly. Additionally, there had to be at least a 30 percentage point difference in the probabilities of success between the two levels. The result was a grouping of assessment items by the levels between which they discriminate. These anchor items were then reviewed by subject area experts who, using their knowledge of mathematics and student performance, generalized from the items to descriptions of the types of skills exhibited at each level. Further details of the anchoring process appear in Appendix F.

The National Assessment Governing Board has determined that achievement levels shall be the first and primary way of reporting NAEP results. Setting achievement levels is a method for setting standards on the NAEP assessment that identifies what students should know and be able to do at various points on the mathematics composite. For each grade, three levels were defined—basic, proficient, and advanced. Based on initial policy definitions of these levels, panelists were asked to determine operational descriptions of the levels appropriate with the content and skills assessed in the mathematics assessment. With these descriptions in mind, the panelists were then asked to rate the assessment items in terms of the expected performance of marginally acceptable examinees at each of these three levels. These ratings were then mapped onto the NAEP scale and adjusted downward one standard error of the mean panelist rating to obtain the achievement level cutpoints for reporting. Further details of the achievement level setting process appear in Appendix G.

Chapter 9

DATA ANALYSIS AND SCALING FOR THE 1992 TRIAL STATE ASSESSMENT IN MATHEMATICS

John Mazzeo, Huahua Chang, Edward Kulick, Y. Fai Fong, and Angela Grima

Educational Testing Service

9.1 OVERVIEW

This chapter describes the analyses carried out in the development of the 1992 Trial State Assessment mathematics scales. The procedures used were similar to those employed in the analysis of the 1990 Trial State Assessment (Mazzeo, 1991) and are based on the philosophical and theoretical underpinnings described in the previous chapter. However, the 1990 methods needed to be extended in a number of ways to accommodate the evolving nature of NAEP in general and the Trial State Assessment in particular. The changes incorporated into the 1992 Trial State Assessment included the assessment of both fourth-grade and eighth-grade samples for each jurisdiction, the addition of items measuring estimation skills, and the introduction of extended constructed-response items.

There were five major steps in the analysis of the Trial State Assessment mathematics data, each of which is described in a separate section:

- conventional item and test analyses (section 9.3);
- item response theory (IRT) scaling (section 9.4);
- estimation of state and subgroup proficiency distributions based on the "plausible values" methodology (section 9.5);
- linking of the 1992 Trial State Assessment scales to the corresponding scales from the 1992 national assessment (section 9.6); and
- creation of the Trial State Assessment mathematics composite scale (section 9.7).

To set the context within which to describe the methods and results of scaling procedures, a brief review of the assessment instruments and administration procedures is provided.

9.2 DESCRIPTION OF ITEMS, ASSESSMENT BOOKLETS, AND ADMINISTRATION PROCEDURES

The general design structure of the 1992 Trial State Assessment was the same as that used in 1990. However, the particulars of the 1992 design differed in several respects from those of 1990. First, the 1990 assessment was administered to eighth-grade students only, while the 1992 assessment included samples of both fourth- and eighth-grade public-school students. Second, the 1992 assessment used a somewhat different set of instruments from those used in 1990. The 1992 item pool was based on the same curriculum framework used for 1990 national and Trial State Assessments and contained four blocks of items at each grade level that were identical to blocks administered in 1990. However, the 1992 item pool included an expanded number of blocks containing new material, including a greater proportion of the conventional short constructed-response items and 11 newly developed extended constructed-response items. Each extended constructed-response item required about five minutes to complete and was scored on a 0-to-4 scale. All extended constructed-response items appeared as the last item in their respective blocks. The 1992 item pool also included a block of items measuring estimation skills. This estimation block had been included in the 1990 national assessment but not in the 1990 Trial State Assessment.

The fourth-grade item pool contained 175 items. Of these, 155 were categorized into one of the five content areas: 63 items for Numbers and Operations, 29 items for Measurement, 27 for Geometry, 20 for Data Analysis, Statistics, and Probability, and 16 for Algebra and Functions. These items, consisting of 95 multiple-choice items, 53 short constructed-response items, 5 extended constructed-response items, and 2 "testlets"[1] were divided into 13 mutually exclusive blocks. The composition of each block of items, in terms of content and format, is given in Table 9-1[2]. An additional 20 multiple-choice items measuring estimation abilities were assembled into a separate block of items.

The eighth-grade item pool contained 205 items. One-hundred and eighty-three items, 55 of which were common to the fourth grade, were classified into the five content areas as follows: 58 items for Numbers and Operations, 32 items for Measurement, 36 for Geometry, 28 for Data Analysis, Statistics, and Probability, and 29 for Algebra and Functions. These 183 items, consisting of 116 multiple-choice items, 59 short constructed-response items, 6 extended constructed-response items, and 2 testlets, were divided into 13 mutually exclusive blocks. The composition of each block of items, in terms of content and format, is given in Table 9-2. An additional 22 multiple-choice items measuring estimation abilities were assembled into a separate block of items. Ten of these items were also used at the fourth grade.

Twelve of the 14 fourth-grade blocks contained one or more constructed-response items; one block consisted entirely of constructed-response items. Twelve of the 14 eighth-grade blocks

---

[1] A testlet is a group of items (in the case of NAEP, typically three or four items) that are related to a single content area, topic, or stimulus and are developed and scored as a single unit (see Wainer & Kiely, 1987, for further details and examples of different types of testlets).

[2] The numbers in Tables 9-1 and 9-2 differ slightly from those given in Chapter 2. The numbers in Chapter 2 do not reflect the grouping of certain sets of items into testlets for the purposes of scaling.

Table 9-1

1992 NAEP Mathematics Block Composition by Scale and Item Type*
Grade 4

Block	Numbers & Operations					Measurement					Geometry					Data Analysis, Statistics, and Probability					Algebra & Functions					Estimation	Total				
	1	2	3	4	T	1	2	3	4	T	1	2	3	4	T	1	2	3	4	T	1	2	3	4	T	1	1	2	3	4	T
M3	2	2	0	0	4	4	1	0	0	5	1	0	0	0	1	1	1	0	0	2	1	0	0	0	1	0	9	4	0	0	13
M4	7	0	0	0	7	2	0	0	0	2	2	0	0	0	2	1	0	0	0	1	2	0	0	0	2	0	14	0	0	0	14
M5	6	0	0	0	6	2	2	0	0	4	1	2	0	0	3	1	0	0	0	1	3	0	0	0	3	0	13	4	0	0	17
M6	0	4	0	0	4	0	1	0	0	1	0	3	0	0	3	0	1	0	0	1	0	2	0	0	2	0	0	11	0	0	11
M7	3	1	1	0	5	0	1	0	0	1	1	0	0	0	1	1	1	0	0	2	1	0	0	0	1	0	6	3	1	0	10
M8	8	1	0	0	9	2	0	0	0	2	2	0	0	0	2	1	0	0	0	1	1	0	0	0	1	0	14	1	0	0	15
M9	1	2	0	0	3	2	0	0	1	3	0	0	1	0	1	2	0	0	0	2	1	0	0	0	1	0	6	2	1	1	10
M10	0	0	0	0	0	0	1	0	0	1	0	4	0	0	4	0	1	0	0	1	0	0	0	0	0	0	0	6	0	0	6
M11	3	3	0	0	6	2	0	0	0	2	5	0	0	0	5	1	2	0	0	3	0	0	0	0	0	0	11	5	0	0	16
M12	2	4	0	0	6	1	0	0	0	1	1	0	0	0	1	1	2	0	0	3	0	1	0	0	1	0	5	7	0	0	12
M13	3	1	0	0	4	2	0	0	0	2	0	3	0	0	3	1	1	0	0	2	0	0	1	0	1	0	6	5	1	0	12
M14	3	1	1	0	5	1	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	1	0	1	2	0	5	2	1	1	9
M15	3	1	0	0	4	3	1	0	0	4	0	0	0	0	0	0	0	1	0	1	0	1	0	0	1	0	6	3	1	0	10
M16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	20	0	0	0	20
Total	41	20	2	0	63	21	7	0	1	29	14	12	1	0	27	10	9	1	0	20	9	5	1	1	16	20	105	53	5	2	175

---

* Item types: 1 = Multiple-choice; 2 = Short constructed-response; 3 = Extended constructed-response; 4 = Testlets; T = Total

Table 9-2

1992 NAEP Mathematics Block Composition by Scale and Item Type*
Grade 8

	Numbers & Operations					Measurement					Geometry					Data Analysis, Statistics, and Probability					Algebra & Functions					Estimation	Total				
Block	1	2	3	4	T	1	2	3	4	T	1	2	3	4	T	1	2	3	4	T	1	2	3	4	T	1	1	2	3	4	T
M3	4	0	1	0	5	1	1	0	0	2	0	1	0	0	1	1	0	0	1	2	2	1	0	0	3	0	8	3	1	1	13
M4	7	0	0	0	7	4	0	0	0	4	4	0	0	0	4	2	0	0	0	2	4	0	0	0	4	0	21	0	0	0	21
M5	7	0	0	0	7	2	3	0	0	5	1	2	0	0	3	3	0	0	0	3	3	0	0	0	3	0	16	5	0	0	21
M6	0	3	0	0	3	0	1	0	0	1	0	6	0	0	6	0	3	0	0	3	0	3	0	0	3	0	0	16	0	0	16
M7	3	1	0	0	4	0	1	0	0	1	1	1	1	0	3	1	1	0	1	3	1	1	0	0	2	0	6	5	1	1	13
M8	8	0	0	0	8	2	0	0	0	2	2	0	0	0	2	1	2	0	0	3	3	0	0	0	3	0	16	2	0	0	18
M9	0	3	0	0	3	1	0	0	0	1	1	0	0	0	1	2	0	1	0	3	1	0	0	0	1	0	5	3	1	0	9
M10	0	0	0	0	0	0	2	0	0	2	0	4	0	0	4	0	1	0	0	1	0	0	0	0	0	0	0	7	0	0	7
M11	2	3	0	0	5	2	1	0	0	3	5	0	0	0	5	2	2	0	0	4	2	0	0	0	2	0	13	6	0	0	19
M12	3	0	0	0	3	0	1	0	0	1	1	1	0	0	2	1	0	0	0	1	1	0	1	0	2	0	6	2	1	0	9
M13	2	2	0	0	4	0	1	1	0	2	2	0	0	0	2	1	0	0	0	1	1	1	0	0	2	0	6	4	1	0	11
M14	1	2	1	0	4	2	0	0	0	2	1	0	0	0	1	1	0	0	0	1	1	0	0	0	1	0	6	2	1	0	9
M15	4	1	0	0	5	5	1	0	0	6	2	0	0	0	2	0	1	0	0	1	2	1	0	0	3	0	13	4	0	0	17
M16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22	22	0	0	0	22
Total	41	15	2	0	58	19	12	1	0	32	20	15	1	0	36	15	10	1	2	28	21	7	1	0	29	22	138	59	6	2	205

---

* Item types: 1 = Multiple-choice; 2 = Short constructed-response; 3 = Extended constructed-response; 4 = Testlets; T = Total

also contained one or more constructed-response items; two blocks consisted entirely of constructed-response items. The questions contained in the constructed-response block at fourth grade and one of the two constructed-response blocks at eighth grade required the manipulation of geometric shapes for their solution. Students assigned these blocks were provided a packet containing the necessary shapes during the time period in which they worked on these items. These and all other constructed-response items were scored by specially trained readers, as described in Chapter 5.

At grade 4, 37 items required the use of a calculator for their solutions. These items appeared in three of the blocks (15 items in block M8, 12 in M12, and 10 in M14). At grade 8, 37 calculator items appeared in three blocks (18 items in block M8, 9 in M12, and 10 in M14). Each student assigned a block containing calculator items was given a Texas Instruments calculator (a TI-108 four-function calculator at grade 4, a TI-30 scientific calculator at grade 8) to use while he or she was working on that block. For each calculator item, both fourth- and eighth-grade students were asked to indicate whether they had in fact used the calculator to answer each item in the blocks for which calculators were made available.

One item at grade 4 required the use of a ruler; five items at grade 8 required the use of a protractor/ruler. Students administered the(se) item(s) were provided with the necessary tools for the 15-minute period they worked on the block containing the item(s).

There were a total of 27 assessment booklets for each grade. The block of estimation items was assembled into a single booklet. The 13 non-estimation blocks were used to form 26 different booklets according to a balanced incomplete block (BIB) design (see Chapter 2 for details). Each of these booklets contained three blocks of items, and each block of items appeared in exactly six booklets. To balance possible block position effect, each block appeared twice as the first block of mathematics items, twice as the second block, and, twice as the third block. In addition, the BIB design required that each block of items be paired in a booklet with every other block of items exactly once.

The design of the 1992 state mathematics assessment required that each student be administered two booklets—one of the 26 booklets in the BIB design, followed by the estimation booklet. Within each administration site, all booklets except the estimation booklet were "spiraled" together in a random sequence and distributed to students sequentially, in the order of the students' names on the Student Listing Form (see Chapter 4). As a result of the BIB design and the spiraling of booklets, a considerable degree of balance was achieved in the data collection process. With the exception of the estimation block, each block of items (and, therefore, each item) was administered to randomly equivalent samples of students of approximately equal size (i.e., about 6/26 of the total sample size) within each jurisdiction and across all jurisdictions. In addition, within and across jurisdictions, randomly equivalent samples of approximately equal size received each particular block of items as the first, second, or third block within a booklet. The full sample of students in each jurisdiction was administered the estimation booklet and all students attempted these items after completing one of the BIB booklets.

As described in Chapter 4, a randomly selected half of the administration sessions within each state were observed by Westat-trained quality control monitors. Thus, within and across states, randomly equivalent samples of students received each block of items in a particular position within a booklet under monitored and unmonitored administration conditions. Randomly equivalent samples of student within and across states also were administered the estimation items under monitored and unmonitored administration conditions.

9.3 ITEM ANALYSES

9.3.1 Conventional Item and Test Analyses

Tables 9-3 and 9-4 contain summary statistics for each block of items for the fourth and the eighth grades respectively. Block-level statistics are provided both overall and, for all but the estimation block, by serial position of the block within booklet. To produce these tables, data from all 44 jurisdictions were aggregated and statistics were calculated using rescaled versions of the final sampling weights provided by Westat. The rescaling, carried out within each jurisdiction, constrained the sum of the sampling weights within that jurisdiction to be equal to its sample size. Use of the rescaled weights does nothing to alter the value of statistics calculated separately within each jurisdiction. However, for statistics obtained from samples that combine students from different jurisdictions, use of the rescaled weights results in a roughly equal contribution of each jurisdiction's data to the final value of the estimate. As discussed in Mazzeo (1991), equal contribution of each jurisdiction's data to the results of the IRT scaling was viewed as a desirable outcome and, as described in the scaling section below, these same rescaled weights were used in carrying out that scaling. Hence, the item analysis statistics shown in Tables 9-3 and 9-4 are consistent with the weighting used in scaling.

Tables 9-3 and 9-4 show the number of students assigned each block of items, the average item score, the average biserial correlation, and the proportion of students attempting the last item in that block. The average item score for the block is the average, over items, of the score means for each of the individual items in the block. For binary-scored multiple-choice and constructed-response items, these score means correspond to the proportion of students who correctly answered each item. For the testlets and extended constructed-response items, the score means were calculated as item score mean divided by the maximum number of points possible.

In NAEP analyses (both conventional and IRT-based), a distinction is made between missing responses at the end of each block (i.e., missing responses subsequent to the last item the student answered) and missing responses prior to the last observed response. Missing responses before the last observed response are considered intentional omissions. In calculating the average score for each item, only students classified as having been presented the item were included in the denominator of the statistic. Intentional omissions were treated as incorrect responses. Missing responses at the end of the block are considered "not-reached," and treated as if they had not been presented to the student. The proportion of students attempting the last item of a block (or, equivalently, 1 minus the proportion of student not reaching the last item) is often used as an index of the degree of speededness associated with the administration of that block of items.

Standard practice at ETS is to treat all nonrespondents to the last item as if they had not reached the item. For multiple-choice and standard constructed-response items, the use of such a convention most often produces a reasonable pattern of results in that the proportion reaching

Table 9-3

Descriptive Statistics for Each Block of Items by Position Within Test Booklet and Overall
Grade 4

Statistic	Position	M3	M4[a]	M5[a,b]	M6[a]	M7	M8[a,c]	M9	M10[d]	M11	M12[c]	M13	M14[c]	M15	M16[e]
Unweighted sample size	1	8568	8517	8532	8519	8470	8553	8528	8432	8509	8566	8590	8508	8545	–
	2	8555	8365	8533	8487	8524	8506	8540	8507	8509	8551	8543	8613	8541	–
	3	8628	8649	8551	8609	8419	8351	8309	8480	8475	8585	8562	8502	8591	–
	ALL	25751	25531	25616	25615	25413	25410	25377	25419	25493	25702	25695	25623	25675	110,455
Average item score	1	0.47	0.43	0.43	0.41	0.38	0.57	0.43	0.37	0.47	0.48	0.50	0.43	0.46	–
	2	0.47	0.43	0.42	0.40	0.39	0.58	0.43	0.35	0.47	0.49	0.51	0.44	0.46	–
	3	0.48	0.43	0.42	0.39	0.39	0.56	0.43	0.36	0.48	0.47	0.51	0.43	0.45	–
	ALL	0.48	0.43	0.42	0.40	0.39	0.57	0.43	0.36	0.47	0.48	0.50	0.43	0.46	.56
Average r-biserial	1	0.59	0.51	0.54	0.66	0.57	0.54	0.61	0.82	0.57	0.61	0.65	0.61	0.64	–
	2	0.59	0.51	0.54	0.67	0.57	0.54	0.61	0.82	0.57	0.61	0.65	0.62	0.64	–
	3	0.61	0.52	0.55	0.67	0.58	0.54	0.60	0.84	0.58	0.63	0.65	0.62	0.64	–
	ALL	0.59	0.51	0.54	0.67	0.57	0.54	0.61	0.83	0.57	0.62	0.65	0.61	0.64	.43
Proportion of students attempting last item	1	0.83	0.95	0.71	0.81	0.92	0.82	0.99	0.90	0.91	0.67	0.95	0.91	0.98	–
	2	0.87	0.96	0.79	0.83	0.94	0.85	0.99	0.92	0.92	0.71	0.97	0.95	0.99	–
	3	0.86	0.95	0.81	0.85	0.95	0.91	0.99	0.94	0.95	0.74	0.97	0.94	0.99	–
	ALL	0.85	0.96	0.77	0.83	0.93	0.86	0.99	0.92	0.92	0.71	0.96	0.93	0.99	1.00

---

[a] Trend block administered in the 1990 Trial State Assessment

[b] Required a ruler

[c] Required a four-function calculator

[d] Required geometric shapes

[e] Estimation block—administered using paced-tape procedures to all students as the fourth and final block

Table 9-4

Descriptive Statistics for Each Block of Items by Position Within Test Booklet and Overall
Grade 8

Statistic	Position	M3	M4[a]	M5[a,b]	M6[a]	M7	M8[a,c]	M9	M10[d]	M11	M12[c]	M13	M14[c]	M15	M16[e]
Unweighted sample size	1	8260	8263	8261	8316	8317	8388	8345	8374	8321	8358	8350	8317	8283	-
	2	8301	8208	8273	8247	8285	8308	8352	8347	8358	8369	8365	8376	8314	-
	3	8388	8443	8349	8323	8217	8234	8210	8267	8275	8315	8314	8339	8351	-
	ALL	24949	24914	24914	24886	24819	24907	24907	24988	24952	25029	25029	25032	24948	107,893
Average item score	1	0.48	0.57	0.69	0.56	0.61	0.49	0.34	0.55	0.62	0.40	0.49	0.35	0.55	-
	2	0.50	0.58	0.69	0.56	0.62	0.49	0.33	0.55	0.63	0.41	0.50	0.37	0.56	-
	3	0.50	0.57	0.68	0.57	0.62	0.50	0.33	0.53	0.62	0.40	0.49	0.36	0.56	-
	ALL	0.49	0.57	0.69	0.56	0.62	0.50	0.33	0.55	0.62	0.40	0.49	0.36	0.56	.56
Average r-biserial	1	0.64	0.53	0.67	0.68	0.62	0.59	0.62	0.79	0.64	0.62	0.62	0.55	0.59	-
	2	0.65	0.53	0.67	0.69	0.62	0.58	0.61	0.78	0.63	0.63	0.62	0.56	0.58	-
	3	0.65	0.52	0.68	0.69	0.63	0.60	0.63	0.80	0.64	0.64	0.62	0.56	0.60	-
	ALL	0.65	0.53	0.67	0.68	0.62	0.59	0.62	0.79	0.64	0.63	0.62	0.56	0.59	.47
Proportion of students attempting last item	1	0.96	0.88	0.79	0.92	0.62	0.59	0.62	0.91	0.90	0.99	0.99	0.93	0.84	-
	2	0.96	0.90	0.80	0.91	0.62	0.58	0.61	0.91	0.88	0.99	0.99	0.94	0.82	-
	3	0.97	0.89	0.85	0.93	0.63	0.60	0.63	0.92	0.93	0.98	0.98	0.95	0.83	-
	ALL	0.97	0.89	0.81	0.92	0.62	0.59	0.62	0.92	0.90	0.99	0.99	0.94	0.83	1.00

---

[a] Trend block administered in the 1990 Trial State Assessment

[b] Required a protractor/ruler

[c] Required a scientific calculator

[d] Required geometric shapes

[e] Estimation block—administered using paced-tape procedures to all students as the fourth and final block

the last item is not dramatically smaller than the proportion reaching the next-to-last item. However, for the blocks that ended with extended constructed-response items, use of the standard ETS convention resulted in an extremely large drop in the proportion of students attempting the final item. A drop of such magnitude seemed somewhat implausible. Therefore, for blocks ending with an extended constructed-response items, students who answered the next-to-last item but did not respond to the extended constructed-response item were classified as having intentionally omitted the last item.

The average biserial correlation is the average, over items, of the item-level biserial correlations (r-biserial). For each item-level r-biserial, total block number-correct score (including the item in question, and with students receiving zero points for all not-reached items) was used as the criterion variable for the correlation. Data from students classified as not reaching the item were omitted from the calculation of the statistic.

As is evident from Tables 9-3 and 9-4, the difficulty and the internal consistency of the blocks varied somewhat. Such variability was expected since these blocks were not created to be parallel in either difficulty or content. Based on the proportion of students attempting the last item, 2 blocks for the fourth grade and 3 blocks for the eighth grade seem to be somewhat speeded. Only 77 percent of the fourth-grade students taking block M5 (which required the use of a ruler) and 71 percent taking block M12 (which required a calculator) reached the last item in the block. Only 62 percent of the eighth-grade students taking block M7 (a calculator block), 59 percent taking block M8, and 62 percent taking block M9 reached the last items of these blocks.

These two tables also indicate that there was little variability in average item scores or average biserial correlations for each block by serial position within the assessment booklet. This suggests that serial position within booklet had a negligible effect on the overall difficulty of the block. However, for the fourth grade, one aspect of block level performance that did differ by serial position was the proportion of students attempting the last item in the block. As shown in Table 9-3, for blocks M5, M6, M7, M8, M10, M11, M12,and M14, the percentage of the students attempting the last item increased as the serial position of the block increased. Perhaps fourth-grade students are able to work more quickly in later blocks or are better able to pace themselves as a result of their experience with the first block of items that they attempt. It is interesting to note that this effect was particularly salient for blocks M5, M8, and M12. Blocks M8 and M12 required the use of a calculator and block M5 required the use of a ruler. Only one block at grade 8 showed a substantial position effect. Interestingly, this was again block M5, which required the use of a protractor/ruler.

As mentioned earlier, in an attempt to maintain rigorous standardized administration procedures across the states, a randomly selected 50 percent of all sessions within each state was observed by a Westat-trained quality control monitor. Observations from this random half of the sessions provided information about the quality of administration procedures and the frequency of departures from standardized procedures in the monitored sessions (see Chapter 4, section 4.3.6 for a discussion of the results of these observations). In addition, unexpectedly large differences in results from monitored and unmonitored sessions (i.e., differences larger than those to be expected due to sampling fluctuation) provided a means to identify instances of cheating, breaches of test security, or other breaks in standardization occurring in the unmonitored sessions that might threaten the validity of assessment results.

When results were aggregated over all participating jurisdictions, there was little difference between the performance of students who attended monitored or unmonitored sessions. The average item score (over all 14 blocks and over all 44 participating jurisdictions) for the fourth-grade students was .46 for both monitored and unmonitored sessions. The average item score for the eighth grade for both monitored and unmonitored sessions was .52. Tables 9-5 and Table 9-6 provide, for each block of items, the average item score, average r-biserial, and the proportion of students attempting the last item for students whose sessions were monitored and students whose sessions were not monitored. Little or no differences in average item performance by session type were evident. These aggregate results are quite consistent with those observed in the 1990 Trial State Assessment, where no evidence was found that students who attended monitored sessions performed differently than those who attended unmonitored sessions.

Figure 9-1 presents stem-and-leaf displays for grades 4 and 8 of the differences between monitored and unmonitored average item scores (over all 14 blocks) for each of the 44 jurisdictions participating in the 1992 Trial State Assessment. Stem-and-leaf displays, developed by Tukey (1977), are somewhat like histograms. For this figure (and all other stem-and-leaf displays that follow), the first column contains observation depths (Hoaglin, Mosteller, & Tukey, 1983). Depths are essentially cumulative frequencies, counted up from the lowest value for score intervals ("stems") below the median and counted down from the highest value for score intervals above the median. The second column contains a count of the number of "leaves" on each stem. In histogram terms, these counts would be considered frequencies. The remainder of the figure contains the stem-and-leaf display. The combination of a stem with each of its leaves gives the actual value of one observation (i.e., the difference in average item scores for monitored and unmonitored sessions in a participating jurisdiction).

At the fourth grade, the median difference (monitored minus unmonitored) was .0015. For 21 jurisdictions, the difference was negative (i.e., students from unmonitored sessions scored higher than students from monitored sessions), with the largest difference being -.021. For the remaining 23 jurisdictions, the difference was positive, with the largest difference being .028. In evaluating the magnitude of these differences, it should be noted that the standard error for a difference in proportions from independent simple random samples of size 1,250 (half the typical total state sample size of 2,500) from a population with a true proportion of .5 is about .02. For samples with complex sampling designs like NAEP, the standard errors tend to be larger than those associated with simple random sampling. A reasonable estimate of the design effect for average item scores based on past NAEP experience with item proportion correct statistics is about 1.5 (Johnson & Rust, 1992), which suggests that a typical estimate of the standard error of the difference between monitored and unmonitored sessions would be about .024. For 41 of the 44 participants, the absolute differences in item score means at fourth grade were less than .02, and all but one were less than .024. In summary, differences in results obtained from the two types of sessions at the fourth grade were well within the bounds expected due to sampling fluctuation.

At the eighth grade, the median difference was essentially zero (.0005). However, the distribution of differences was somewhat negatively skewed. For 19 jurisdictions, the differences were negative. Two were larger in absolute magnitude than .024, both of which were negative (-.037 for the Virgin Islands and -.033 for Florida). For the remaining 25 jurisdictions, the difference was zero or positive, and all of them were less than or equal to .015 in magnitude.

Table 9-5

Block-level Descriptive Statistics for Monitored and Unmonitored Sessions
Grade 4

Statistic	M3	M4^{a}	M5^{a,b}	M6^{a}	M7	M8^{a,c}	M9	M10^{d}	M11	M12^{c}	M13	M14^{c}	M15	M16^{e}
Unweighted sample size														
Unmonitored	12891	12740	12777	12796	12668	12695	12706	12695	12738	12813	12857	12799	12764	55069
Monitored	12860	12791	12839	12819	12745	12715	12671	12727	12755	12889	12838	12824	12911	55386
Average item score														
Unmonitored	0.48	0.43	0.42	0.40	0.39	0.57	0.43	0.36	0.47	0.48	0.50	0.43	0.46	0.56
Monitored	0.48	0.43	0.43	0.40	0.39	0.57	0.44	0.36	0.47	0.48	0.51	0.43	0.46	0.56
Average r-biserial														
Unmonitored	0.59	0.51	0.54	0.67	0.57	0.54	0.61	0.83	0.57	0.62	0.65	0.61	0.64	0.43
Monitored	0.60	0.52	0.55	0.67	0.58	0.54	0.61	0.82	0.58	0.62	0.65	0.62	0.64	0.43
Proportion of students attempting last item														
Unmonitored	0.85	0.95	0.77	0.82	0.93	0.85	0.99	0.91	0.92	0.70	0.96	0.93	0.98	1.00
Monitored	0.86	0.96	0.77	0.83	0.93	0.86	0.99	0.92	0.93	0.71	0.97	0.93	0.99	1.00

---

[a] Trend block administered in the 1990 Trial State Assessment

[b] Required a ruler

[c] Required a four-function calculator

[d] Required geometric shapes

[e] Estimation block—administered using paced-tape procedures to all students as the fourth and final block

Table 9-6

Block-level Descriptive Statistics for Monitored and Unmonitored Sessions
Grade 8

Statistic	M3	M4[a]	M5[a,b]	M6[a]	M7	M8[a,c]	M9	M10[d]	M11	M12[c]	M13	M14[c]	M15	M16[e]
Unweighted sample size														
Unmonitored	12601	12569	12593	12587	12535	12566	12542	12551	12568	12610	12609	12636	12625	54398
Monitored	12348	12345	12290	12299	12284	12364	12365	12437	12384	12432	12420	12396	12323	53495
Average item score														
Unmonitored	0.49	0.57	0.68	0.56	0.62	0.50	0.33	0.55	0.62	0.41	0.50	0.36	0.56	0.56
Monitored	0.49	0.58	0.69	0.56	0.61	0.49	0.33	0.55	0.62	0.40	0.50	0.36	0.56	0.56
Average r-biserial														
Unmonitored	0.65	0.53	0.67	0.68	0.62	0.59	0.62	0.79	0.64	0.64	0.62	0.56	0.59	0.47
Monitored	0.65	0.53	0.67	0.69	0.62	0.59	0.62	0.79	0.64	0.63	0.62	0.56	0.59	0.47
Proportion of students attempting last item														
Unmonitored	0.97	0.90	0.82	0.92	0.98	0.67	0.96	0.92	0.90	0.98	0.99	0.94	0.83	.100
Monitored	0.97	0.88	0.81	0.92	0.98	0.65	0.96	0.92	0.90	0.99	0.99	0.94	0.83	.100

---

[a] Trend block administered in the 1990 Trial State Assessment

[b] Required a protractor/ruler

[c] Required a scientific calculator

[d] Required geometric shapes

[e] Estimation block—administered using paced-tape procedures to all students as the fourth and final block

Figure 9-1

Stem-and-leaf Display* of State-by-state Differences in Average Item Scores (Monitored - Unmonitored)

GRADE 4 ITEM POOL

N = 44, Median = 0.0015, Quartiles = -0.0045, 0.0100
Decimal point is 2 places to the left of the colon

```
 1     1  -2 : 1
 4     3  -1 : 875
 5     1  -1 : 4
11     6  -0 : 976665
21    10  -0 : 4433332111
       6   0 : 122344
17     5   0 : 78899
12     6   1 : 001234
 6     4   1 : 5688
 2     1   2 : 0
 1     1   2 : 8
```

GRADE 8 ITEM POOL

N = 44, Median = 0.0005, Quartiles = -0.0105, 0.0070
Decimal point is 2 places to the left of the colon

```
 2     2  -3 : 73
 4     2  -2 : 30
13     9  -1 : 855443100
19     6  -0 : 988743
      17   0 : 00012344555667778
 8     8   1 : 12233345
```

---

* The first column of numbers shows observation depths; the second column shows the number of observations; the remainder of the figure contains the stem-and-leaf display.

Although the presence of somewhat larger differences at grade 8 is worth noting, even these larger differences are probably less than 2 standard errors in magnitude. Thus, in sum, differences in results obtained from the two types of sessions at the eighth grade were also within the bounds expected due to sampling fluctuation.

9.3.2 Differential Item Functioning (DIF) Analyses

Prior to scaling, differential item functioning (DIF) analyses were carried out on 1992 NAEP mathematics data from the national cross-sectional samples at grades 4, 8, and 12 and the Trial State Assessment samples at grades 4 and 8. The purpose of these analyses was to identify items that were differentially difficult for various subgroups and to reexamine such items with respect to their fairness and their appropriateness for inclusion in the scaling process. The information in this section focuses mainly on the analyses conducted on the Trial State Assessment data. A description of the results based on the national assessment will appear in the forthcoming technical report for that assessment.

The DIF analyses were based on the Mantel-Haenszel chi-square procedure, as adapted by Holland and Thayer (1988). The procedure tests the statistical hypothesis that the odds of correctly answering an item are the same for two groups of examinees that have been matched on some measure of proficiency (usually referred to as the matching criterion). The groups being compared are often referred to as the focal group (usually a minority group of interest, such as Black examinees or female examinees) and the reference group (usually White examinees or male examinees). The measure of proficiency used is typically the number-correct score on some collection of items. Separate analyses were performed for each block of items (i.e., data were pooled across booklets containing the block being analyzed), and number correct score on the block of items in question was used as the measure of proficiency.

For each item in the assessment, an estimate was produced of the Mantel-Haenszel common odds-ratio, expressed on the ETS delta scale for item difficulty . The estimates indicate the difference between reference group and focal group item difficulties (measured in ETS delta scale units), and typically run between about +3 and -3. Positive values indicate items that are differentially easier for the focal group than the reference group after making and adjustment for the overall level of proficiency in the two groups. Similarly, negative values indicate items that are differentially harder for the focal group than the reference group. It is common practice at ETS to categorize each item into one of three categories (Petersen, 1988): "A" (items exhibiting no DIF), "B" (items exhibiting a weak indication of DIF), or "C" (items exhibiting a strong indication of DIF). Items in category A have Mantel-Haenszel values that do not differ significantly from 0 at the alpha = .05 level. Two conditions must be met in order for items to fall in category B. The Mantel-Haenszel value for the item must: (1) be significantly greater than 0 but not significantly greater than 1 at the .05 level, and, (2) must be less than 1.5 in absolute magnitude. Category C items are those with Mantel-Haenszel values that are significantly greater than 1 and larger than 1.5 in absolute magnitude.

For each block of items at each grade a single set of analyses was carried out based on equal-sized random samples of data from all participating jurisdiction. Each set of analyses involved four reference group/focal group comparisons: male/female, White/Asian American, White/Black, and White/Hispanic. The first subgroup in each comparison is the reference group; the second subgroup is the focal group.

All analyses used rescaled sampling weights. A separate rescaled weight was defined for each comparison as:

$$\textit{Rescaled Weight} = \textit{Original Weight} \times \frac{\textit{Total Sample Size}}{\textit{Sum of the Weights}}$$

where the total sample size is the total number of students for the two groups being analyzed (e.g., for the White/Hispanic comparison, the total number of White and Hispanic examinees in the sample at that grade), and the sum of the weights is the sum of the sampling weights of all the students in the sample for the two groups being analyzed. Four rescaled weights were computed for White examinees—one for the gender comparison and three for the race/ethnicity comparisons. Two rescaled overall weights were computed for the Asian American, Black, and Hispanic examinees—one for the gender comparison and another for the appropriate race/ethnicity comparison.

The ETS generalized program IANA83 was used to carry out the DIF analyses. Two-sided modification[3] was used. In the calculation of number-correct scores for the matching criterion, both not-reached and omitted items were considered as wrong responses. For each item, calculation of the Mantel-Haenszel statistic did not include data from examinees who did not reach the item in question. Because the Mantel-Haenszel procedure, as currently implemented, is appropriate only for dichotomously scored items, the extended constructed-response items had to be scored dichotomously for the DIF analyses. Extended constructed responses rated as "satisfactory" or "extended" were scored as correct; all other responses were scored as incorrect.

At grade 4, 159 items were analyzed; at grade 8, 211 items were analyzed[4]. Items common to both grades underwent separate DIF analyses for each grade. Tables 9-7 and 9-8 provide a summary of the results of the DIF analyses for the grade 4 and grade 8 collections of items grouped by content or skill area. For each grade, the tables provide six sets of five frequency distributions for the categorized Mantel-Haenszel statistics for the items in each of the scales. The leftmost frequency distribution gives the number (and percent) of items in each of five categories (C+, B+, A, B-, C-) based on the largest absolute DIF value obtained for the item across the four reference group/focal group comparisons that were carried out. The remaining four frequency distributions give the number of items with indices in each DIF category for each of the four reference group/focal group comparisons.

---

[3] Modification refers to the procedure in which items classified as "C" items in an initial DIF analysis are deleted from the matching criterion, and a second DIF analysis is run. Two-sided means that "C" items are deleted from the criterion, regardless of which group they favor.

[4] Separate DIF indices were calculated for the individual component items of each testlet. No additional DIF analyses were carried out on the overall testlet score.

Table 9-7
Frequency Distributions of DIF Statistics for Grade 4 Items Grouped by Content or Skill Area

Category of Maximum Absolute DIF Value For All Comparisons			Number of Items in Category of DIF Value for Each Comparison (Reference Group/Focal Group)			
DIF Category*	Number	Percent	Male/Female	White/Black	White/Hispanic	White/Asian Amer.
Numbers and Operations						
C+	0	0.0	0	0	0	0
B+	11	17.5	1	1	7	4
A	43	68.3	62	58	51	53
B-	7	11.1	0	4	3	6
C-	2	3.2	0	0	2	0
Measurement						
C+	0	0.0	0	0	0	0
B+	4	13.8	0	0	1	3
A	15	51.7	26	25	24	20
B-	6	20.7	3	2	2	3
C-	4	13.8	0	2	2	3
Geometry						
C+	1	3.7	1	0	0	0
B+	8	29.6	2	3	0	4
A	13	48.1	23	22	25	21
B-	2	7.4	1	0	2	1
C-	3	11.1	0	2	0	1
Data Analysis, Statistics, and Probability						
C+	2	10.0	1	0	0	1
B+	0	0.0	0	0	0	1
A	11	55.0	18	14	17	14
B-	2	10.0	1	4	2	0
C-	5	25.0	0	2	1	4
Algebra and Functions						
C+	0	0.0	1	0	0	0
B+	2	11.8	1	1	0	0
A	11	64.7	14	14	16	15
B-	3	17.6	1	2	0	1
C-	1	5.9	0	0	1	1
Estimation						
C+	1	5.0	0	1	0	0
B+	3	15.0	0	2	0	1
A	12	60.0	19	13	20	19
B-	3	15.0	1	3	0	0
C-	1	5.0	0	1	0	0

---

* Categories are A, B, and C. (+) indicates items in the category that are differentially easier for the focal group; (-) indicates items in the category that are differentially more difficult for the focal group.

Table 9-8
Frequency Distributions of DIF Statistics for Grade 8 Items Grouped by Content or Skill Area

Category of Maximum Absolute DIF Value For All Comparisons			Number of Items in Category of DIF Value for Each Comparison (Reference Group/Focal Group)			
DIF Category*	Number	Percent	Male/Female	White/Black	White/Hispanic	White/Asian Amer.
Numbers and Operations						
C+	6	10.3	1	1	1	4
B+	10	17.2	7	6	1	5
A	25	43.1	46	46	49	43
B-	13	22.4	4	2	6	6
C-	4	6.9	0	3	1	0
Measurement						
C+	0	0.0	0	0	0	0
B+	4	12.5	1	1	1	1
A	17	53.1	29	26	28	26
B-	7	21.9	2	4	2	3
C-	4	12.5	0	1	1	2
Geometry						
C+	3	8.3	0	1	0	2
B+	10	27.8	2	3	6	6
A	14	38.9	30	29	28	27
B-	9	25.0	4	3	2	1
C-	0	0.0	0	0	0	0
Data Analysis, Statistics, and Probability						
C+	1	3.6	0	0	1	0
B+	3	10.7	1	2	0	1
A	11	39.3	26	23	23	15
B-	6	21.4	1	1	3	7
C-	7	25.0	0	2	1	5
Algebra and Functions						
C+	5	17.2	1	0	0	4
B+	9	31.0	2	1	0	7
A	10	34.5	26	24	27	17
B-	3	10.3	0	3	1	1
C-	2	6.9	0	1	1	0
Estimation						
C+	0	0.0	0	0	0	0
B+	2	9.1	0	1	0	1
A	17	77.3	22	18	22	20
B-	2	9.1	0	2	0	1
C-	1	4.5	0	1	0	0

---

[*] Categories are A, B, and C. (+) indicates items in the category that are differentially easier for the focal group; (-) indicates items in the category that are differentially more difficult for the focal group.

A total of 20 items were classified as "C" items for at least one of the analyses for the fourth-grade Trial State Assessment data; 33 items were classified as "C" items for at least one of the analyses for the eighth-grade Trial State Assessment data. For the grade 4 items, 80 percent of the "C" items (16 out of 20) were differentially more difficult for at least one of the four focal groups (female, Black, Hispanic, or Asian American examinees). Nine of these items covered topics in the content areas of Measurement and Data Analysis, Statistics, and Probability. The grade 8 "C" items were split about equally between those favoring the reference group and those favoring the focal group. A relatively large proportion of items (12 of 28) covering topics in Data Analysis, Statistics, and Probability were differentially more difficult (B- or C-) for Asian American examinees. In contrast, differentially functioning items covering topics in Algebra and Functions, and, to a lesser extent, Geometry, tended to favor Asian American examinees.

Following standard practice at ETS for DIF analyses conducted on final test forms, all "C" items were reviewed by a committee of trained test developers and subject-matter specialists. Such committees are charged with making judgments about whether or not the differential difficulty of an item is *unfairly* related to group membership. As pointed out by Zieky (1993):

> It is important to realize that *DIF* is not a synonym for *bias*. The item response theory based methods, as well as the Mantel-Haenszel and standardization methods of DIF detection, will identify questions that are not measuring the same dimension(s) as the bulk of the items in the matching criterion....Therefore, judgement is required to determine whether or not the difference in difficulty shown by a DIF index is *unfairly* related to group membership. The judgement of fairness is based on whether or not the difference in difficulty is believed to be related to the construct being measured....The fairness of an item depends directly on the purpose for which a test is being used. For example, a science item that is differentially difficult for women may be judged to be fair in a test designed for certification of science teachers because the item measures a topic that every entry-level science teacher should know. However, that same item, with the same DIF value, may be judged to be unfair in a test of general knowledge designed for all entry-level teachers. (p. 340)

The committee assembled to review NAEP items included both ETS staff and outside members with expertise in the field. It was the committee's judgment that none of the "C" items for the national or Trial State Assessment data were functioning differentially due to factors irrelevant to test objectives. Hence, none of the items were removed from scaling due to differential item functioning.

9.4 ITEM RESPONSE THEORY (IRT) SCALING

Items at each grade were sorted into six distinct sets, one for each of the five mathematics content areas and one for estimation. Figure 9-2 contains stem-and-leaf displays of the average scores for the items comprising each of the six fourth-grade sets. Figure 9-3 contains corresponding results for the eighth-grade item sets. The averages are based on the entire sample of students in the Trial State Assessment and use the same rescaled sampling

Figure 9-2

Stem-and-leaf Display* of Average Scores for Items, by Scale, for Grade 4

NUMBERS AND OPERATIONS

N = 63, Median = 0.45, Quartiles = 0.3, 0.65
Decimal point is 1 place to the left of the colon

```
  1    1   0 : 9
  6    5   1 : 07889
 15    9   2 : 000246678
 24    9   3 : 002556667
      11   4 : 02333455677
 28    9   5 : 024457789
 19    9   6 : 011578899
 10    5   7 : 01444
  5    4   8 : 3889
  1    1   9 : 1
```

MEASUREMENT

N = 29, Median = 0.43, Quartiles = 0.33, 0.6
Decimal point is 1 place to the left of the colon

```
  1    1   0 : 5
  2    1   1 : 9
  6    4   2 : 0112
 12    6   3 : 035788
       7   4 : 1333459
 10    2   5 : 67
  8    2   6 : 03
  6    4   7 : 2368
  2    2   8 : 78
```

(continued)

---

* The first column of numbers shows observation depths; the second column shows the number of observations; the remainder of the figure contains the stem-and-leaf display.

Figure 9-2 (continued)

Stem-and-leaf Display* of Average Scores for Items, by Scale, for Grade 4

GEOMETRY

N = 27, Median = 0.4, Quartiles = 0.25, 0.65
Decimal point is 1 place to the left of the colon

```
  1    1    0 : 6
  4    3    1 : 134
  9    5    2 : 02569
 13    4    3 : 2247
       3    4 : 013
 11    4    5 : 2249
  7    4    6 : 5578
  3    1    7 : 7
  2    0    8 :
  2    2    9 : 01
```

DATA ANALYSIS, STATISTICS, AND PROBABILITY

N = 20, Median = 0.475, Quartiles = 0.27, 0.545
Decimal point is 1 place to the left of the colon

```
  1    1    1 : 1
  6    5    2 : 13668
  7    1    3 : 2
       5    4 : 14788
  8    6    5 : 233677
  2    2    6 : 34
```

(continued)

---

* The first column of numbers shows observation depths; the second column shows the number of observations; the remainder of the figure contains the stem-and-leaf display.

Figure 9-2 (continued)

Stem-and-leaf Display* of Average Scores for Items, by Scale, for Grade 4

ALGEBRA AND FUNCTIONS

N = 16, Median = 0.34, Quartiles = 0.27, 0.52
Decimal point is 1 place to the left of the colon

```
1    1    1 : 8
5    4    2 : 0559
     5    3 : 01267
6    1    4 : 9
5    3    5 : 047
2    1    6 : 4
1    0    7 :
1    0    8 :
1    1    9 : 0
```

ESTIMATION

N = 19, Median = 0.54, Quartiles = 0.43, 0.68
Decimal point is 1 place to the left of the colon

```
1    1    2 : 5
4    3    3 : 458
7    3    4 : 358
     5    5 : 01446
7    3    6 : 068
4    1    7 : 2
3    3    8 : 366
```

---

* The first column of numbers shows observation depths; the second column shows the number of observations; the remainder of the figure contains the stem-and-leaf display.

Figure 9-3

Stem-and-leaf Display* of Average Scores for Items, by Scale, for Grade 8

NUMBERS AND OPERATIONS

N = 58, Median = 0.635, Quartiles = 0.42, 0.8
Decimal point is 1 place to the left of the colon

```
 1     1    0 : 8
 2     1    1 : 6
 7     5    2 : 11488
12     5    3 : 01389
18     6    4 : 002699
24     6    5 : 012579
       9    6 : 001334699
25     9    7 : 001256679
16    10    8 : 0011233559
 6     6    9 : 011233
```

MEASUREMENT

N = 32, Median = 0.565, Quartiles = 0.265, 0.73
Decimal point is 1 place to the left of the colon

```
 2     2    0 : 88
 6     4    1 : 0289
 9     3    2 : 267
13     4    3 : 2235
14     1    4 : 7
       3    5 : 058
15     5    6 : 11448
10     5    7 : 23399
 5     5    8 : 03479
```

(continued)

---

* The first column of numbers shows observation depths; the second column shows the number of observations; the remainder of the figure contains the stem-and-leaf display.

Figure 9-3 (continued)

Stem-and-leaf Display* of Average Scores for Items, by Scale, for Grade 8

GEOMETRY

N = 36, Median = 0.5, Quartiles = 0.315, 0.695
Decimal point is 1 place to the left of the colon

```
 1    1    0 : 8
 1    0    1 :
 7    6    2 : 144999
13    6    3 : 012469
18    5    4 : 24589
18    5    5 : 13789
13    4    6 : 0089
 9    3    7 : 057
 6    5    8 : 02268
 1    1    9 : 2
```

DATA ANALYSIS, STATISTICS, AND PROBABILITY

N = 28, Median = 0.495, Quartiles = 0.215, 0.69
Decimal point is 1 place to the left of the colon

```
 1    1    0 : 9
 5    4    1 : 1256
 9    4    2 : 0129
11    2    3 : 56
14    3    4 : 568
14    3    5 : 189
11    4    6 : 1335
 7    4    7 : 3467
 3    3    8 : 799
```

(continued)

---

* The first column of numbers shows observation depths; the second column shows the number of observations; the remainder of the figure contains the stem-and-leaf display.

Figure 9-3 (continued)

Stem-and-leaf Display* of Average Scores for Items, by Scale, for Grade 8

ALGEBRA AND FUNCTIONS

N = 29, Median = 0.48, Quartiles = 0.32, 0.69
Decimal point is 1 place to the left of the colon

```
 1    1   0 : 8
 1    0   1 :
 6    5   2 : 00678
10    4   3 : 1256
      5   4 : 12468
14    5   5 : 00146
 9    4   6 : 5999
 5    2   7 : 58
 3    1   8 : 0
 2    2   9 : 56
```

ESTIMATION

N = 21, Median = 0.6, Quartiles = 0.41, 0.64
Decimal point is 1 place to the left of the colon

```
 2    2   2 : 58
 5    3   3 : 147
 8    3   4 : 168
10    2   5 : 49
      7   6 : 0034449
 4    2   7 : 35
 2    2   8 : 66
```

---

* The first column of numbers shows observation depths; the second column shows the number of observations; the remainder of the figure contains the stem-and-leaf display.

weights described in the previous section. As a whole, both the fourth- and the eighth-grade students found the set of Algebra and Functions items to be the most difficult. The fourth graders found the set of Estimation items the easiest, while the eighth-grade students found the set of Numbers and Operations items to be easiest.

Separate IRT-based scales corresponding to each of the item sets defined above were developed using the scaling models described in Chapter 8. For each grade, six scales were produced by separately calibrating the sets of items classified in each of the five content areas and the items in the estimation block. Since there were two grades and each had six scales, a total of 12 distinct calibrations were carried out.

For the reasons discussed in Mazzeo (1991), for each scale at each grade, a single set of item parameters for each item was estimated and used for all jurisdictions. Item parameter estimation was carried out using a 25 percent systematic random sample of the students participating in the 1992 Trial State Assessment at each grade and included equal numbers of students from each participating jurisdiction, half from monitored sessions and half from unmonitored sessions. For the fourth-grade calibrations, the sample consisted of 27,720 students, with 630 students being sampled from each of the 44 participating jurisdictions. For the eighth-grade calibrations, the sample consisted of 27,016 students, with 614 students being sampled from each jurisdiction. As was done for 1990, all calibrations were carried out using the rescaled sampling weights described earlier in an effort to ensure that each jurisdiction's data contributed equally to the determination of the item parameter estimates.

As mentioned above, the sample used for item calibration was also constrained to contain an equal number of students from the monitored and unmonitored sessions from each of the participating jurisdictions. To the extent that items may have functioned differently in monitored and unmonitored sessions, the single set of item parameter obtained define a sort of average item characteristic curve for the two types of sessions. Tables 9-5 and 9-6 (shown earlier) presented block-level item statistics that suggested little, if any, differences in item functioning by session type. Figures 9-4 and 9-5 present the results of supplementary analyses organized by scale.

Figures 9-4 (for grade 4) and 9-5 (for grade 8) contain plots of differences in score means (monitored minus unmonitored) against the score means for the monitored sessions for the items in each of the six scales. At grade 4, the differences between session type appear small on all scales, with a slight tendency for performance to be higher in the monitored sessions. At grade 8, for all but one scale, the average scores were quite similar for the two types of sessions with a tendency for performance to be slightly higher in the unmonitored sessions. For the eighth-grade Estimation scale, however, average score means were consistently higher in the unmonitored sessions than in the monitored sessions for almost all of items. Although the average difference over all items is small (.002), the item-by-item consistency of the results suggests that departures from standardized administration procedures in the unmonitored sessions may have occurred in one or more of the jurisdictions.

Figure 9-4

Differences in Average Item Scores (Monitored Minus Unmonitored) Plotted Against Monitored Average Item Scores, Grade 4

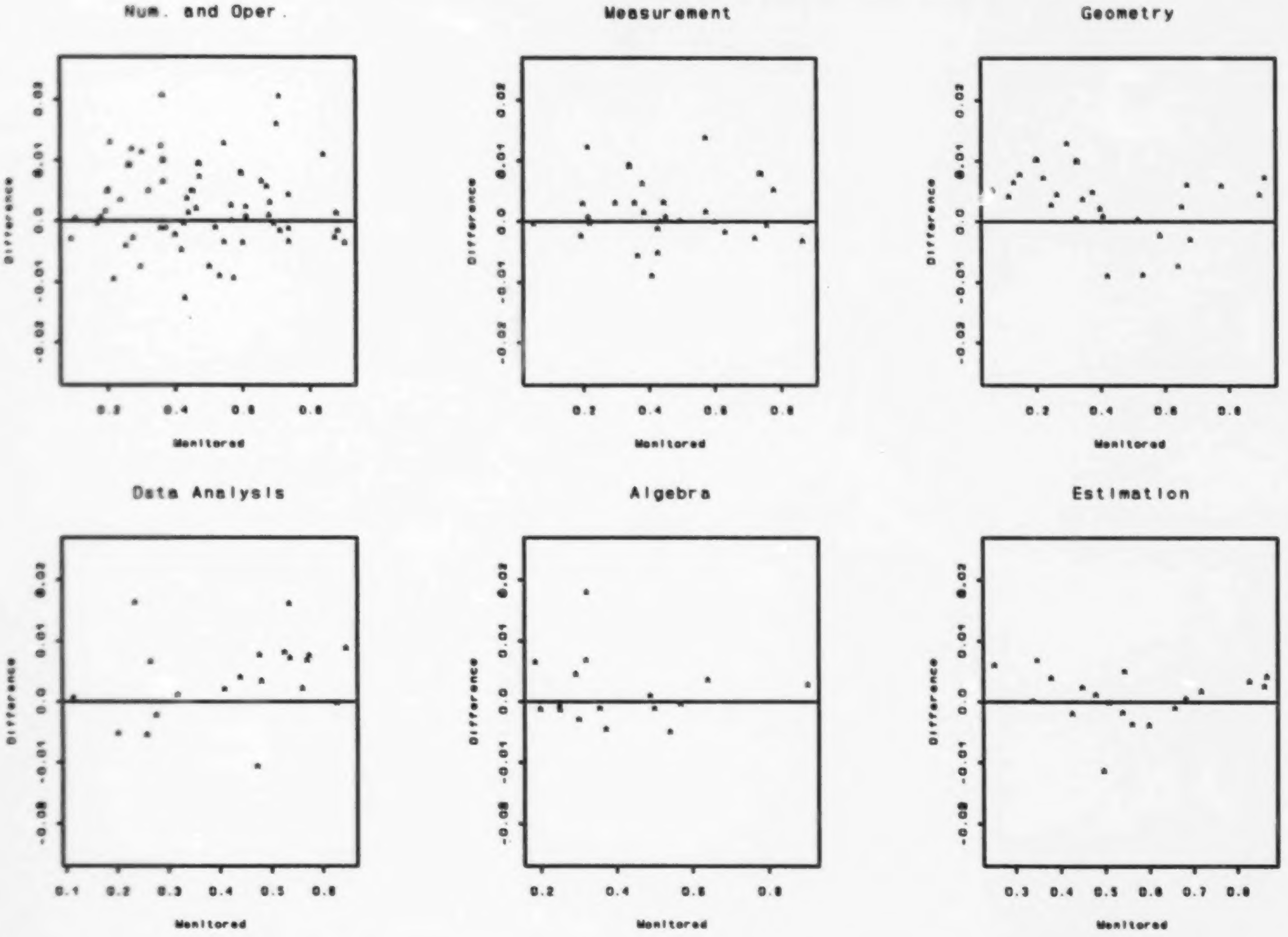


Figure 9-5

Differences in Average Item Scores (Monitored Minus Unmonitored)
Plotted Against Monitored Average Item Scores, Grade 8

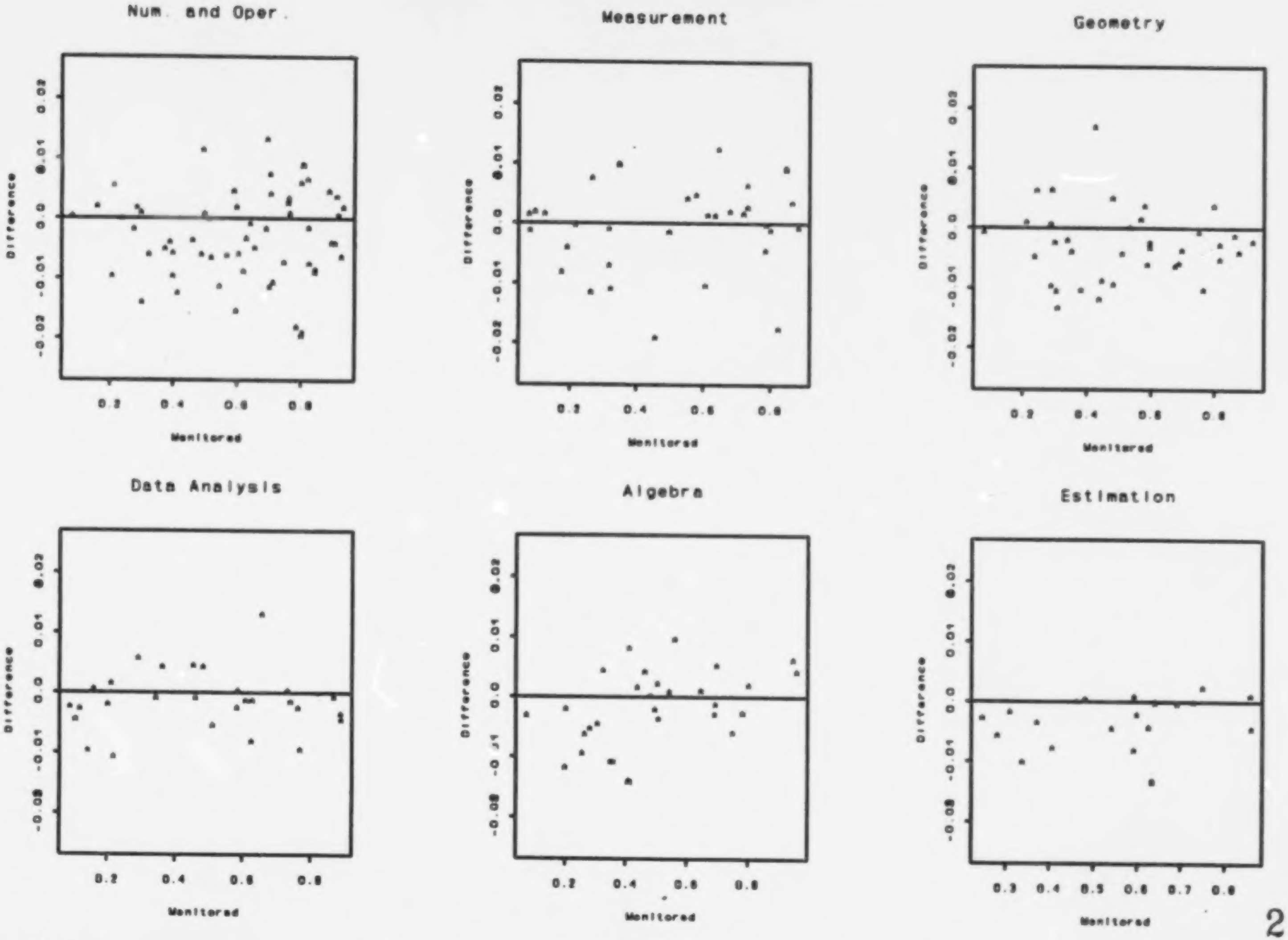


Figure 9-6 contains a stem-and-leaf display of the differences between monitored and unmonitored sessions averaged over items in the eighth-grade Estimation scale for each of the participating jurisdictions. The average item scores were not uniformly higher for all jurisdictions and the median difference was essentially zero (-.0019). In addition, the magnitude of the differences are all within a reasonable range given the expected variation due to sampling. However, there does appear to be a slight tendency toward higher performance by unmonitored sessions on the estimation items. For 25 of the 44 jurisdictions, the differences were negative; for 14, the mean for the unmonitored sessions exceeded that of the monitored sessions by more than .01. A difference of this magnitude in the reverse direction occurred in only 5 jurisdictions. One of these differences, -.031 in Florida, is somewhat large compared with the magnitude of the differences observed in the other jurisdictions. The reasons for this general tendency toward higher performance by unmonitored sessions—and in particular on the somewhat larger difference observed in Florida—cannot be determined.

9.4.1 Item Parameter Estimation

For each grade and each subscale, item parameter estimates were obtained by the NAEP BILOG/PARSCALE program, which combines Mislevy and Bock's (1982) BILOG and Muraki and Bock's (1991) PARSCALE computer programs. The program uses marginal estimation procedures to estimate the parameters of the one-, two-, and three-parameter logistic models, and the generalized partial credit model described by Muraki (1992).

All multiple-choice items were dichotomously scored and were scaled using the three-parameter logistic model. Omitted responses to multiple-choice items were treated as fractionally correct, with the fraction being set to 1 over the number of response options. All short constructed-response items were dichotomously scored and were scaled using the two-parameter logistic model. Omitted responses to short constructed-response items were treated as incorrect.

A key assumption associated with IRT scales is that of conditional independence. Conditional on proficiency, examinee item responses are assumed to be independent. When sets of items are logically dependent on each other, or are based on a single stimulus, this assumption can be violated to a degree that results in aberrant scaling results. In order to avoid possible problems with interitem dependencies, 4 testlets (2 at each grade) were created by combining examinee responses to sets of related items into a single score for each set. At grade 4, one 3-item and one 2-item testlet were created; at grade 8 two 4-item testlets were created. The testlets, rather than their original constituent items, were used in scaling the 1992 mathematics assessment. In all cases, examinee testlet scores were defined as the number of correct responses given to each testlet's constituent items. Examinees omitting all constituents of the testlet were placed in the "zero correct" category of the testlet. Examinees classified as "not reaching" all constituent parts were treated as having not been presented the testlet. All testlets were scaled using the generalized partial credit model.

Figure 9-6

Stem-and-leaf Display* of State-by-state Differences in Average Item Score (Monitored - Unmonitored) for the Grade 8 Estimation Item Pool

N = 44, Median = -0.0019, Quartiles = -0.012, 0.005
Decimal point is 2 places to the left of the colon

```
   1        1   -3  : 1
   1        0   -2  :
   4        3   -2  : 100
   9        5   -1  : 97775
  14        5   -1  : 32210
  17        3   -0  : 875
[illegible] 8   -0  : 44432221
  19        6    0  : 011123
  13        8    0  : 55567889
   5        3    1  : 033
   2        2    1  : 68
```

---

* The first column of numbers shows observation depths; the second column shows the number of observations; the remainder of the figure contains the stem-and-leaf display.

There was a total of 11 extended constructed-response items at grades 4 and 8. Each of these items was also scaled using the generalized partial credit model. Five scoring levels were defined:

0	Wrong, off-task, or omitted
1	Minimal response
2	Partially correct
3	Satisfactory response
4	Elaborated response

Table 9-9 provides a listing of the blocks, content area classifications, and NAEP identification numbers for all extended constructed-response items included in the 1992 assessment.

Table 9-9
Extended Constructed-response Items*
1992 Trial State Assessment in Mathematics

Grade	Block	Scale	NAEP ID
Grade 4	M7	Numbers and Operations	M045401
	M9	Geometry	M041201
	M13	Algebra and Functions	M043501
	M14	Numbers and Operations	M044401
	M15	Data Analysis, Stat., & Prob.	M049001
Grade 8	M3	Numbers and Operations	M051101
	M7	Geometry	M045901
	M9	Data Analysis, Stat., & Prob.	M053101
	M12	Algebra and Functions	M054301
	M13	Measurement	M052201
	M14	Numbers and Operations	M055501

---

* These items always appeared last in their respective blocks. The number of items in each block is shown in Tables 9-1 and 9-2.

Bayes modal-estimates of all item parameters were obtained from the BILOG/PARSCALE program. Prior distributions were imposed on item parameters with the following starting values: thresholds (normal [0,2]); slopes (log-normal [0,.5]); and asymptotes (two-parameter beta with parameter values determined as functions of the number of response options for an item and a weight factor of 50). The locations (but not the dispersions) were updated at each program estimation cycle in accordance with provisional estimates of the item parameters.

As was done for the 1990 Trial State Assessment, item parameter estimation proceeded in two phases. First, the subject ability distribution was assumed fixed (normal [0,1]) and a stable solution was obtained. The parameter estimates from this solution were then used as starting values for a subsequent set of runs in which the subject ability distribution was freed and estimated concurrently with item parameter estimates. After each estimation cycle, the subject ability distribution was re-standardized to have a mean of zero and standard deviation of one. Correspondingly, parameter estimates for that cycle were also linearly re-standardized. During the concurrent estimation phase, convergence problems were encountered for the grade 4 Data Analysis, Statistics, and Probability scale. Therefore, for this scale, the converged normal solution results were used.

During and subsequent to item parameter estimation, evaluations of the fit of the IRT models were carried out for each of the items in the grade 4 and grade 8 item pools. These evaluations were conducted to determine the final composition of the item pool making up the scales by identifying misfitting items that could not be included. Evaluations of model fit were based primarily on a graphical analysis. For binary-scored items, model fit was evaluated by examining plots of estimates of the expected conditional (on θ) probability of a correct response that do not assume a two-parameter or three-parameter logistic model versus the probability predicted by the estimated item characteristic curve (see Mislevy & Sheehan, 1987, p. 302). For the testlets and extended constructed-response items, similar plots were produced for each item category characteristic curve.

As with most procedures that involve evaluating plots of data versus model predictions, a certain degree of subjectivity is involved in determining the degree of fit necessary to justify use of the model. There are a number of reasons why evaluation of model fit relied primarily on analyses of plots rather than seemingly more objective procedures based on goodness-of-fit indices such as the "pseudo chi-squares" produced in BILOG (Mislevy & Bock, 1982). First, the exact sampling distributions of these indices when the model fits are not well understood, even for fairly long tests. Mislevy and Stocking (1987) point out that the usefulness of these indices appears particularly limited in situations like NAEP where examinees have been administered relatively short tests. Work in progress by Stone, Mislevy, and Mazzeo using simulated data suggests that the correct reference chi-square distributions for these indices have considerably fewer degrees of freedom than the value indicated by the BILOG/PARSCALE program and require additional adjustments of scale. However, it is not yet clear how to estimate the correct number of degrees of freedom and necessary scale factor adjustment factors. Consequently, pseudo chi-square goodness-of-fit indices are used only as rough guides in interpreting the severity of model departures.

Second, as discussed in Chapter 8, it is almost certainly the case that, for most items, item-response models hold only to a certain degree of approximation. Given the large samples sizes used in NAEP and the Trial State Assessment, there will be sets of items for which one is almost certain to reject the hypothesis that the model fits the data even though departures are minimal in nature or involve kinds of misfit unlikely to impact on important model-based inferences. In practice, one is almost always forced to temper statistical decisions with judgments about the severity of model misfit and the potential impact of such misfit on final results.

In making decisions about excluding items from the final scales, a balance was sought between being too stringent, hence deleting too many items and possibly damaging the content representativeness of the pool of scaled items, and too lenient, hence including items with model fit poor enough to invalidate the types of model-based inferences made from NAEP results. Items that clearly did not fit the model were not included in the final scales; however, a certain degree of misfit was tolerated for a number of items included in the final scales.

For the large majority of the grade 4 and grade 8 items, the fit of the model was extremely good. Figure 9-7 provides a typical example of what the plots look like for this class of items. The plots that are shown are for items from the grade 8 Algebra and Functions scale. The item at the top of the plot is a binary-scored constructed-response item; the item at the bottom of the plot is a multiple-choice item. In each plot, the y-axis indicates the probability of a correct response and the x-axis indicates proficiency level (theta). The circles show estimates of the conditional (on theta) probability of a correct response that do not assume a logistic form (referred to subsequently as nonlogistic-based estimates). The sizes of the circles are proportional to the estimated density of the theta distribution at the indicated value. The solid line shows the estimated item response function. The item response function provides estimates of the conditional probability of a correct response based on an assumed logistic form. The vertical dashed line indicates the estimated location parameter (b) for the item and the horizontal dashed line (bottom plot only) indicates the estimated lower asymptote (c). Also shown in the plot are the actual values of the item parameter estimates (lower right-hand corner) as well as the proportion of students that answered the item correctly (upper left-hand corner). As is evident from the plots, the nonlogistic-based estimates of conditional probabilities are in extremely close agreement with those given by the estimated item response function.

Figure 9-8 provides an example of a plot for a five-category extended constructed-response item exhibiting good model fit. Like the plots for the binary items, this plot shows two estimates of each item category characteristic curve, one set that does not assume the partial credit model (shown as circles) and one that does (the solid lines). The dashed horizontal lines show the location of the estimated category thresholds for the item (d_1 to d_4; see Chapter 8, sections 8.3.1). The estimates for all parameters for the item in question are also indicated on the plot. As with Figure 9-7, the two sets of estimates agree quite well, although there is a slight tendency for the nonlogistic-based estimates for category two to be somewhat higher than the model-based estimates for theta values greater than 1. An aspect of Figure 9-8 worth noting is the large proportion of examinees that responded in the two lowest response categories for this item[5]. Such results were typical for the extended constructed-response items at both grades. Substantial proportions of examinees were either unable or unwilling to provide even minimally adequate answers to such items.

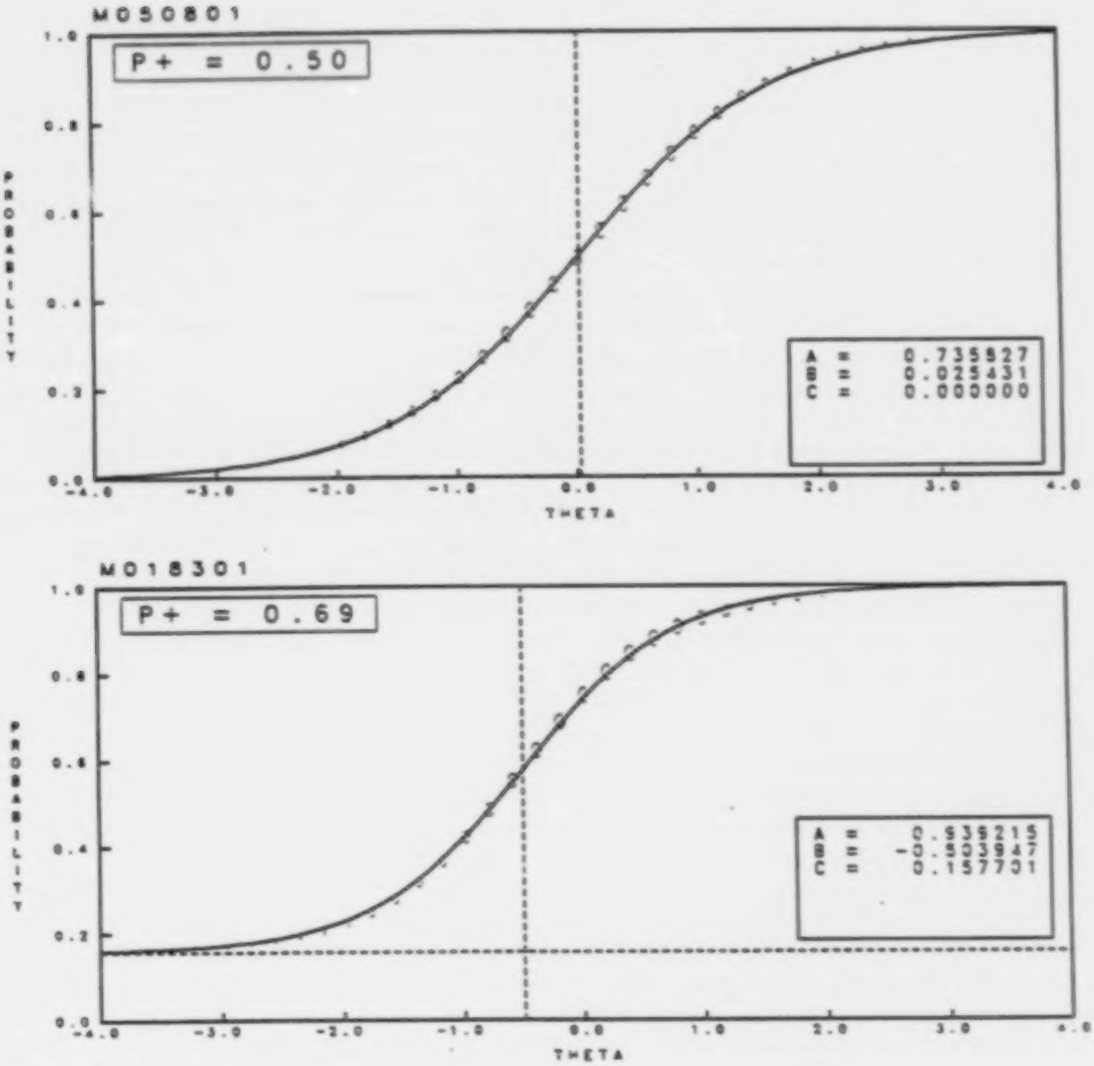
As discussed above, some of the items retained for the final scales display some degree of model misfit. Figures 9-9 (binary-scored items) and 9-10 (extended constructed-response item) provide typical examples of such items. In general, good agreement between nonlogistic and logistic estimates of conditional probabilities were found for the regions of the theta scale

---

[5] This is evidenced by the relatively large size of the circles indicating estimated conditional probabilities for these two categories.

Figure 9-7

Plots* Comparing Empirical and Model-based Estimates of Item Response Functions for Binary-scored Items Exhibiting Good Model Fit

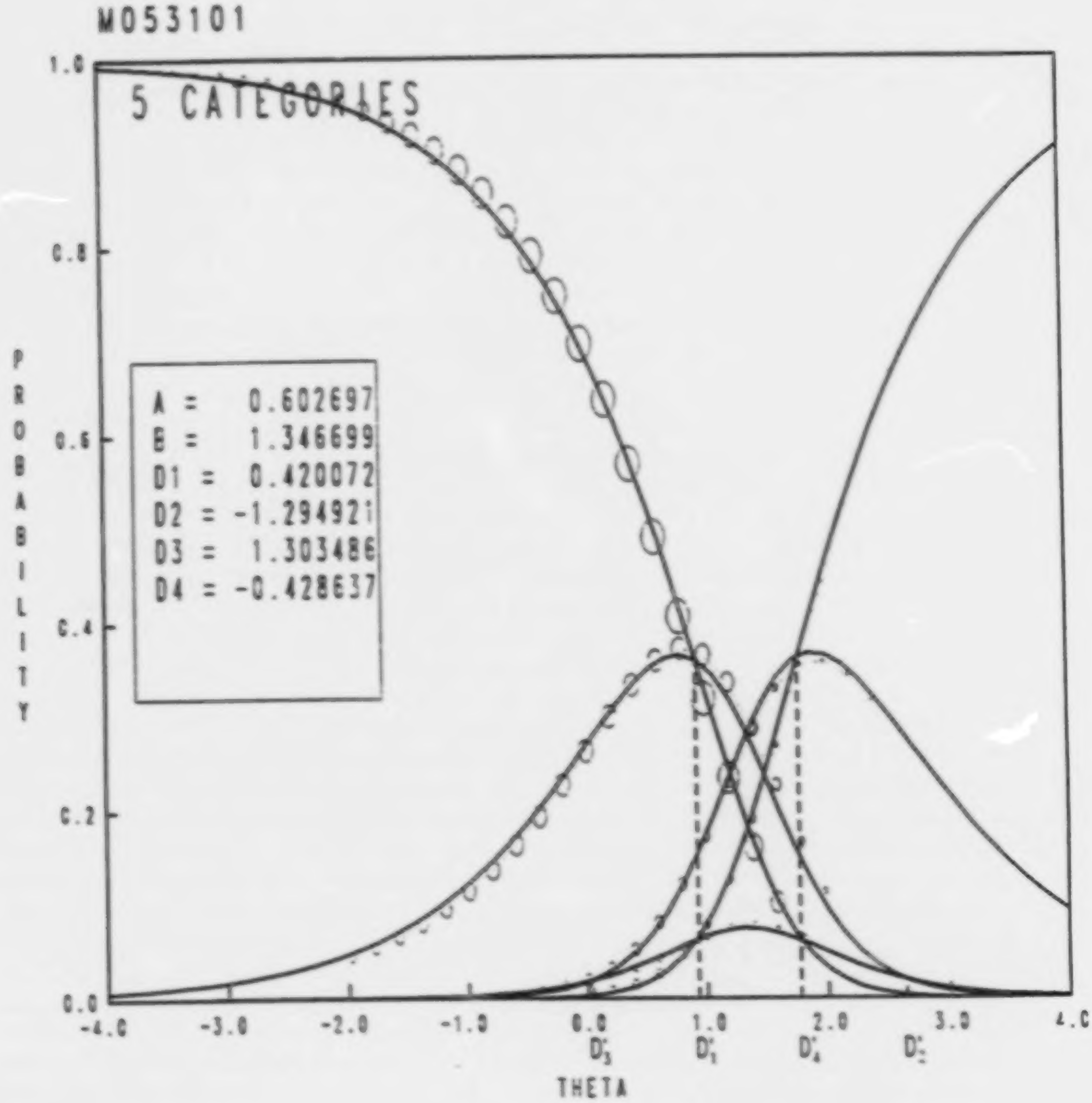


---

* Circles indicate estimated conditional probabilities obtained without assuming a logistic form; solid line indicates estimated item response function assuming a logistic form.

Figure 9-8

Plot* Comparing Empirical and Model-based Estimates of Item Category Characteristic Curves for a Polytomously Scored Item Exhibiting Good Model Fit

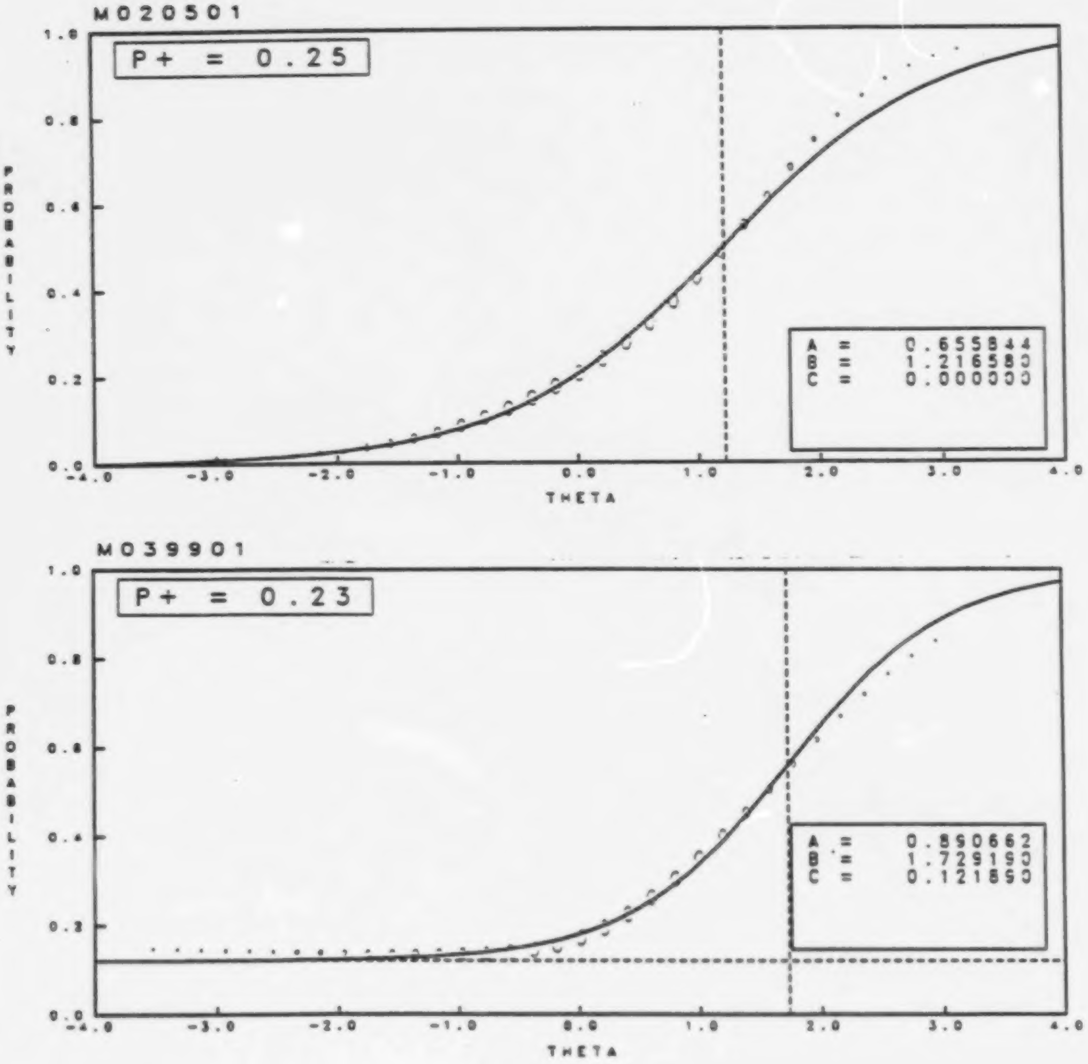


---

* Circles indicate estimated conditional probabilities obtained without assuming a logistic form; solid line indicates estimated item response function assuming a logistic form.

Figure 9-9

Plots* Comparing Empirical and Model-based Estimates of Item Response Functions for Binary-scored Items Exhibiting Some Model Misfit

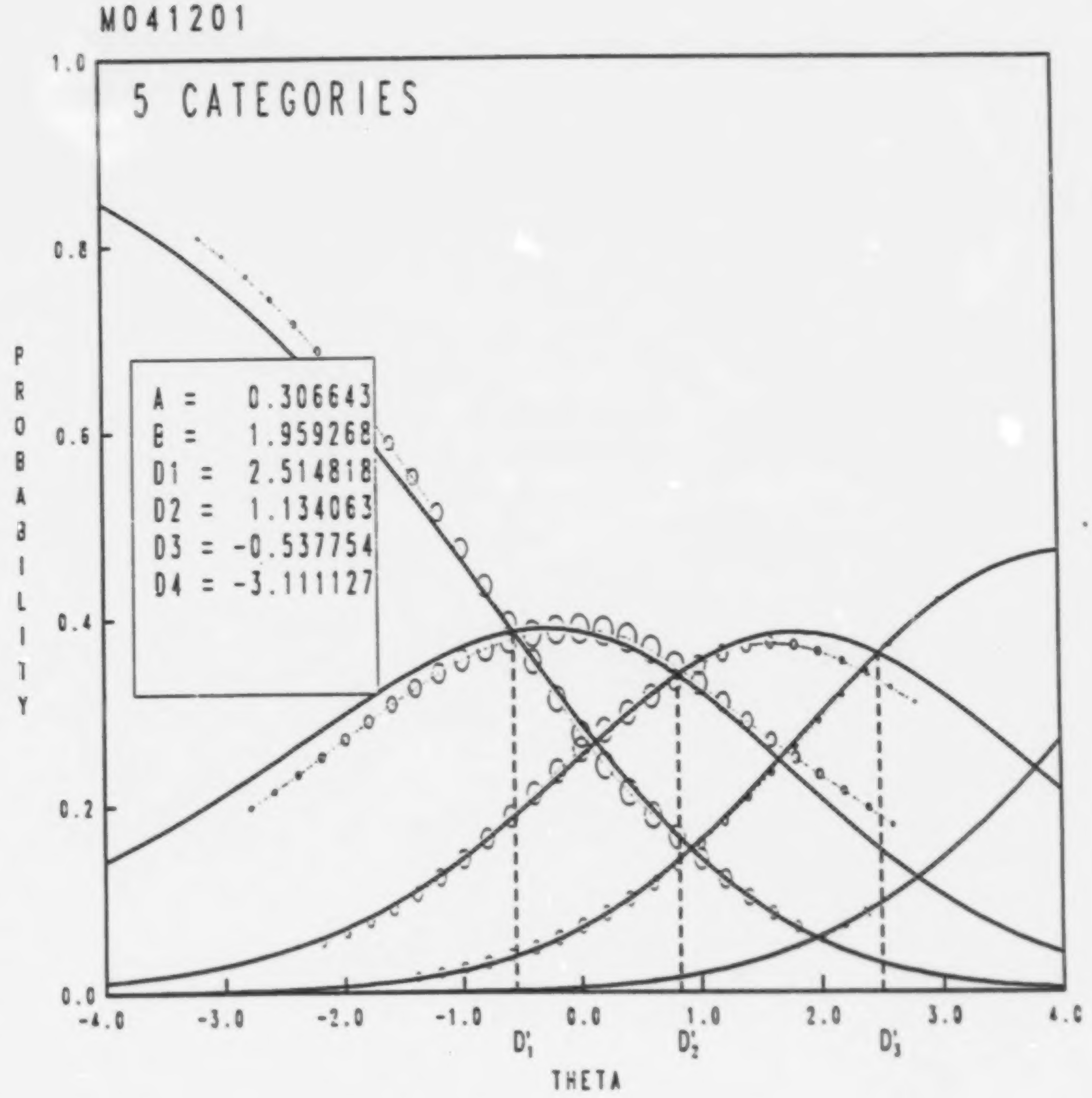


---

* Circles indicate estimated conditional probabilities obtained without assuming a logistic form; solid line indicates estimated item response function assuming a logistic form.

Figure 9-10

Plot* Comparing Empirical and Model-based Estimates of Item Category Characteristic Curves for a Polytomously Scored Item Exhibiting Some Model Misfit



---

* Circles indicate estimated conditional probabilities obtained without assuming a logistic form; solid line indicates estimated item response function assuming a logistic form.

with theta values in the tails of the subject ability distributions. As noted in Chapter 5, two of the extended constructed-response items, one at grade 4 (see Figure 9-10) and one at grade 8, had interreader reliabilities somewhat lower than those of the remaining items. Both of these items did exhibit some degree of model misfit. However, the primary effect of lower interreader reliability is to increase the imprecision of measurement rather than to bias the results.

Only two of the administered items (one from the grade 4 estimation block and one from the grade 8 estimation block) were not included in the final scales. Plots for these items are given in Figure 9-11. As is evident from the nonlogistic-based estimates in the plots, both items appear to have nonmonotonic item characteristic curves. Students with higher levels of proficiency exhibit lower chances of success than do students with lower proficiency. Logistic-based estimates of conditional probabilities are, by definition, monotonically increasing. Hence, the model does not fit.

Table 9-10 lists the items that received special treatment during the scaling process. Included in the table are the block locations and item numbers for the items that were combined into testlets as well as for those that were excluded from the final scales. These items received identical special treatment in the production of the 1992 national scales. No other items in either assessment received special treatment. The IRT parameters for the items included in the Trial State Assessment are listed in Appendix D.

Table 9-10
Items from the 1992 Trial State Assessment in Mathematics Receiving Special Treatment

NAEP ID	Grade	Block	Order in Block	Content Area	Treatment	Reason for Treatment
M040401 M040402 M040403	4	M9	2a 2b 2c	Measurement	Combined to form M040461	Local dependencies across items
M044201 M044202	4	M14	7 8	Algebra	Combined to form M044261	Local dependencies across items
M050201 M050202 M050203 M050204	8	M3	4a 4b 4c 4d	Data Analysis, Statistics, and Probability	Combined to form M050261	Local dependencies across items
M045801 M045802 M045803 M045804	8	M7	12a 12b 12c 12d	Data Analysis, Statistics, and Probability	Combined to form M045861	Local dependencies across items
M032101	4,8	M16	2	Estimation	Not scaled - grade 8 only	Nonmonotonic IRF in 90 & 92
M032801	4,8	M16	9	Estimation	Not scaled - grade 4 only	Nonmonotonic IRF in 90 & 92

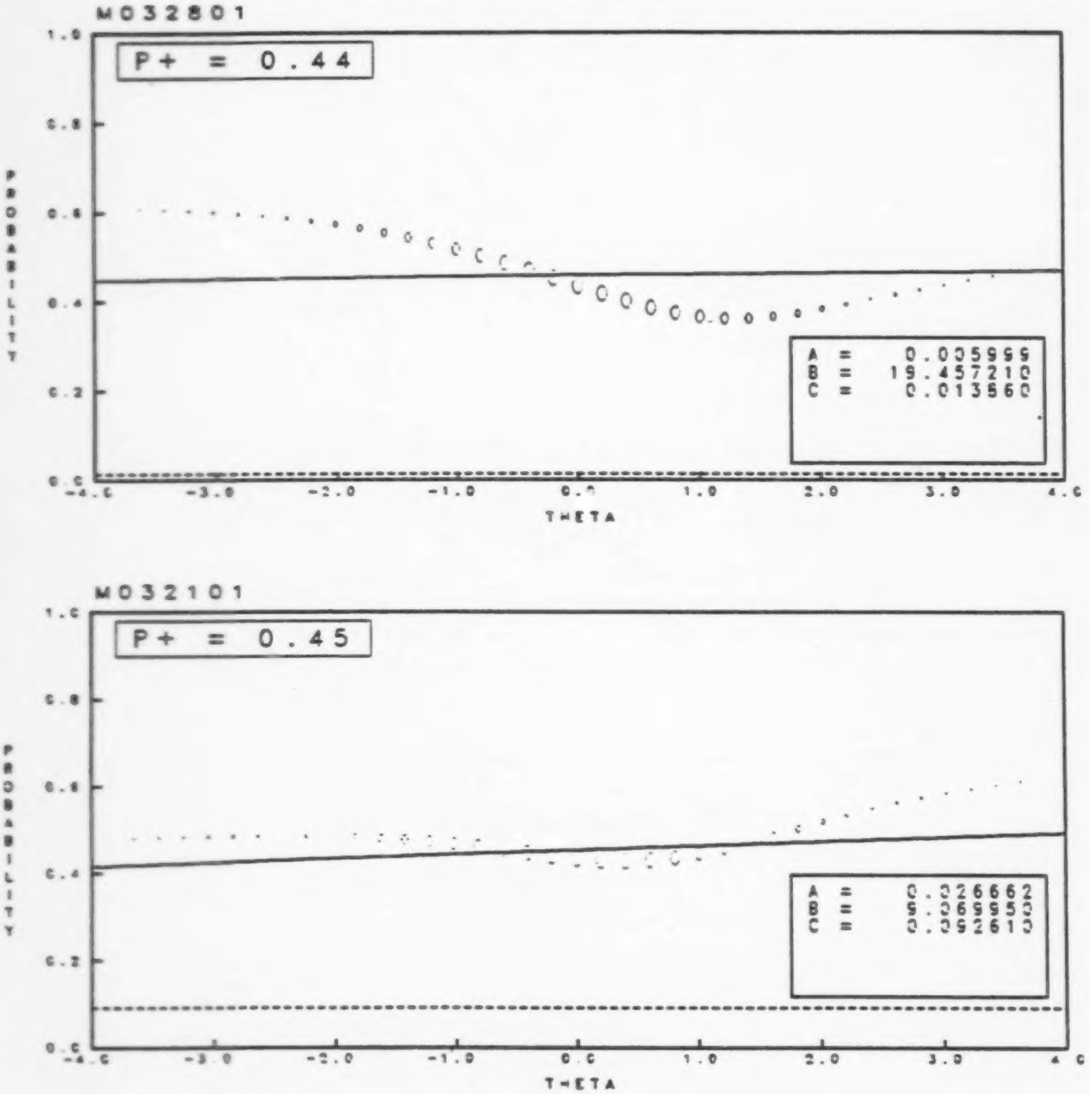
9.5 ESTIMATION OF STATE AND SUBGROUP PROFICIENCY DISTRIBUTIONS

The proficiency distributions in each state (and for important subgroups within each state) were estimated by using the multivariate plausible values methodology and the

BEST COPY AVAILABLE

Figure 9-11

Plots* Comparing Empirical and Model-based Estimates of Item Response Functions for Items Dropped from Scaling Due to Model Misfit



---

* Circles indicate estimated conditional probabilities obtained without assuming a logistic form; solid line indicates estimated item response function assuming a logistic form.

corresponding MGROUP computer program (described in Chapter 8; see also Mislevy, 1991). The MGROUP program (Sheehan, 1985; Rogers, 1991), which was originally based on the procedures described by Mislevy and Sheehan (1987), was used in the 1990 Trial State Assessment of mathematics. The 1992 Trial State Assessment used an enhanced version of MGROUP, based on modifications described by Thomas (1992), to estimate the proficiency distribution for both the fourth and the eighth grades in each state. As described in the previous chapter, MGROUP estimates proficiency distributions using information from student's item responses, student background variables, and the item parameter estimates obtained from the BILOG/PARSCALE program.

The enhancements included in the 1992 version of MGROUP included the replacement of Monte Carlo integration by analytic calculations, new methods for computing student-level posterior means and variances, and the generation of Γ values from their posterior distributions for the imputation of student proficiency values. Simulation studies indicate that the enhanced MGROUP produces more accurate estimates of subscale variances and correlations (Thomas, 1992) than did the previous versions of MGROUP.

For the reasons discussed in Mazzeo (1991), separate conditioning models were estimated at each grade for each jurisdiction. This resulted in the estimation of 88 distinct conditioning models. At each grade, the background variables included in each jurisdiction's model (denoted *y* in Chapter 8) were principal component scores derived from the within-state correlation matrix of selected main-effects and two-way interactions associated with a wide range of student, teacher, school, and community variables. A set of five multivariate plausible values was drawn for each student who participated in the Trial State Assessment.

As was the case in 1990, plans for reporting each jurisdiction's results required analyses examining the relationships between proficiencies and a large number of background variables. The background variables included student demographic characteristics (e.g., the race/ethnicity of the student, highest level of education attained by parents), students' perceptions about mathematics, student behavior both in and out of school (e.g., amount of TV watched daily, amount of mathematics homework done each day), the type of mathematics class being taken (e.g., algebra or general fourth- or eighth-grade mathematics), the amount of emphasis on various topics included in the assessment provided by the students' teachers, and a variety of other aspects of the students' background and preparation, the background and preparation of their teachers, and the educational, social, and financial environment of the schools they attended.

As described in the previous chapter, to avoid biases in reporting results and to minimize biases in secondary analyses, it is desirable to incorporate measures of a large number of independent variables in the conditioning model. When expressed in terms of contrast-coded main effects and interactions, the number of variables to be included totaled 258 at grade 4 and 303 at grade 8. Appendix C provides a listing of the full set of contrasts defined at each grade. These contrasts were the common starting point in the development of the conditioning models for each of the participating jurisdictions.

Because of the large number of these contrasts and the fact that, within each jurisdiction, some contrasts had zero variance, some involved relatively small numbers of individuals, and some were highly correlated with other contrasts or sets of contrasts, an effort was made to reduce the dimensionality of the predictor variables in each jurisdiction's MGROUP models. As was done for the 1990 Trial State Assessment, the original background variable contrasts were standardized and transformed into a set of linearly independent variables by extracting separate sets of principal components (one set for each grade for each of the 44 jurisdictions) from the within-jurisdiction correlation matrices of the original contrast variables. The principal components, rather than the original variables, were used as the independent variables in the conditioning model. As was done for the 1990 Trial State Assessment, the number of principal components included for each state was the number required to account for approximately 90 percent of the variance in the original contrast variables. Research based on data from the 1990 Trial State Assessment suggests that results obtained using such a subset of the components will differ only slightly from those obtained using the full set (Mazzeo, Johnson, Bowker, & Fong, 1992).

Tables 9-11 (for grade 4) and 9-12 (for grade 8) contain a listing of the number of principal components included in and the proportion of proficiency variance accounted for by the conditioning model for each of the 44 participating jurisdictions. It is important to note that the proportion of variance accounted for by the conditioning model differs across scales within a state, across grades within a state, and across states within a scale. Such variability is not unexpected for at least two reasons. First, there is no reason to expect the strength of the relationship between proficiency and demographics to be identical across all grades and states. In fact, one of the reasons for fitting separate conditioning models is that the strength and nature of this relationship may differ across states. Second, the homogeneity of the demographic profile also differs across states. As with any correlational analysis, the restriction of the range in the predictor variables will attenuate the relationship.

Figures 9-12 (for grade 4) and 9-13 (for grade 8) provide boxplots (Tukey, 1977) of the estimated within-jurisdiction correlations among the six scales. One boxplot is provided for each of the 15 unique scale pairs and each boxplot is based on 44 data points (i.e., the estimates of the indicated correlations from each of the 44 participating jurisdictions). The plotted values, taken directly from the revised MGROUP program, are estimates of the within-jurisdiction correlations *conditional on the set of principal components included in the conditioning model*. The box portion shows the locations of the 25th, 50th, and 75th percentiles. Generally, the whiskers extend to the minimum and maximum values. However, values more than 1.5 interquartile ranges from the median are plotted as individual points.

The number and nature of the scales that were produced were consistent with the recommendations for reporting that were given by the National Assessment Planning Project (see Chapter 2). Reporting results on multiple scales is typically most informative when each of the scales provides unique information about the profile of knowledge and skills possessed by the students being assessed. In such cases, one would hope to see relatively low correlations among the subscales. However, with a couple of exceptions, the correlations among the 1992 mathematics scales are high across all jurisdictions, almost always exceeding .7 and quite often exceeding .9. This is particularly noteworthy when one considers that these are correlations *conditional* on a rather large set of background variables. The *marginal* correlations between subscales would be higher, particularly for those correlations in the .7 to .8 range. In particular, the correlations among three of the scales (Numbers and Operations; Data Analysis, Statistics, and Probability; and Algebra and Functions) are extremely high (rarely falling below .9) at both

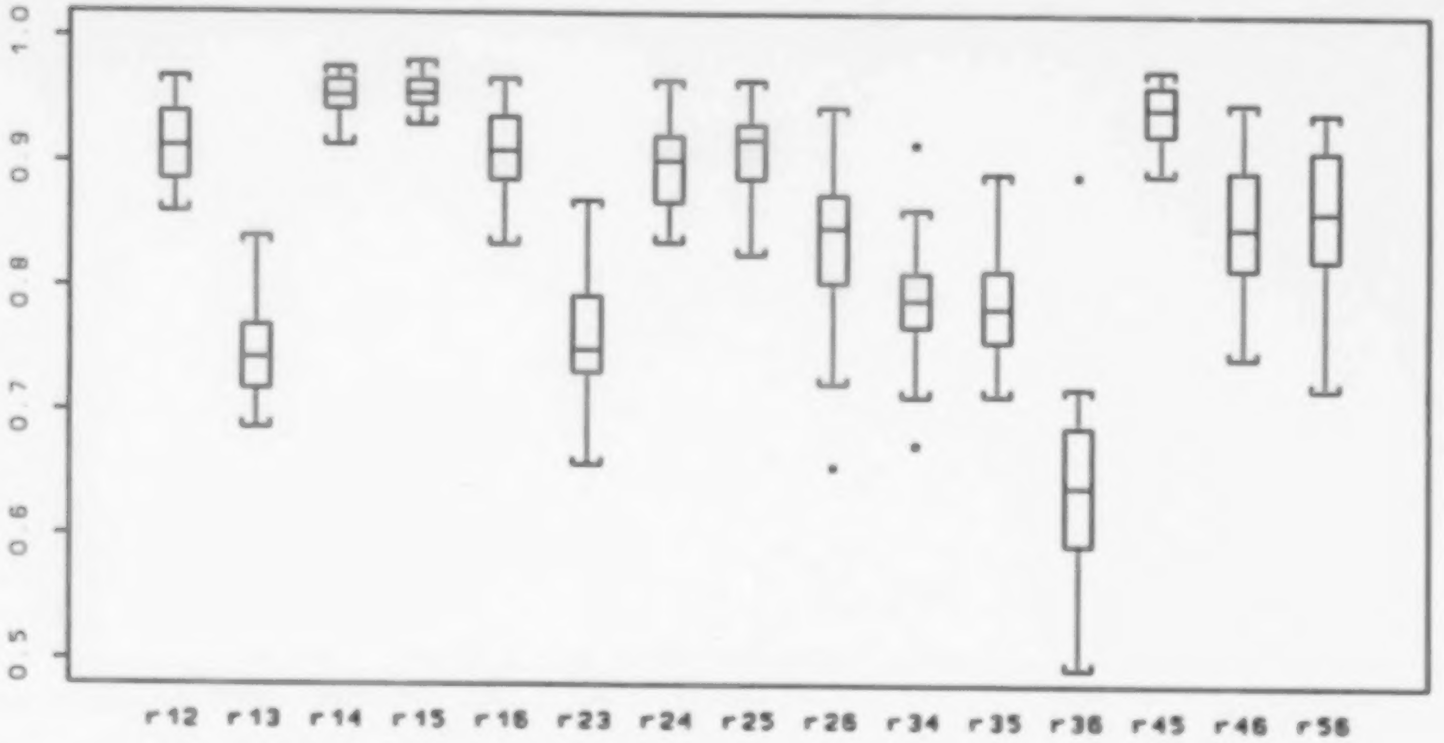
Table 9-11
Proportion of Proficiency Variance Accounted for by Grade 4 Conditioning Models

State	Number of Principal Components	Numbers and Operations	Measurement	Geometry	Data Analysis, Statistics and Probability	Algebra and Functions	Estimation
Alabama	126	0.60	0.66	0.52	0.64	0.66	0.61
Arizona	129	0.57	0.55	0.43	0.58	0.57	0.59
Arkansas	129	0.55	0.59	0.53	0.61	0.57	0.64
California	128	0.59	0.63	0.50	0.67	0.66	0.64
Colorado	128	0.49	0.51	0.41	0.53	0.57	0.58
Connecticut	126	0.59	0.63	0.49	0.64	0.68	0.68
Delaware	122	0.65	0.67	0.57	0.71	0.72	0.67
District of Columbia	128	0.57	0.57	0.51	0.62	0.58	0.63
Florida	131	0.58	0.57	0.45	0.62	0.63	0.64
Georgia	129	0.66	0.67	0.55	0.68	0.68	0.66
Guam	106	0.66	0.64	0.64	0.73	0.65	0.65
Hawaii	129	0.55	0.56	0.47	0.58	0.61	0.53
Idaho	127	0.35	0.30	0.13	0.44	0.31	0.58
Indiana	124	0.52	0.51	0.32	0.51	0.59	0.63
Iowa	123	0.56	0.57	0.55	0.59	0.63	0.63
Kentucky	131	0.57	0.56	0.32	0.62	0.60	0.56
Louisiana	130	0.59	0.64	0.62	0.67	0.65	0.67
Maine	122	0.50	0.56	0.39	0.60	0.64	0.55
Maryland	129	0.69	0.67	0.70	0.70	0.62	0.78
Massachusetts	127	0.50	0.61	0.38	0.59	0.59	0.58
Michigan	123	0.67	0.62	0.62	0.62	0.55	0.71
Minnesota	120	0.61	0.61	0.52	0.66	0.66	0.60
Mississippi	129	0.55	0.52	0.46	0.68	0.59	0.63
Missouri	126	0.52	0.63	0.40	0.48	0.54	0.61
Nebraska	124	0.55	0.61	0.51	0.66	0.69	0.56
New Hampshire	126	0.51	0.49	0.29	0.59	0.53	0.54
New Jersey	125	0.60	0.66	0.58	0.56	0.61	0.68
New Mexico	125	0.64	0.54	0.50	0.62	0.56	0.64
New York	126	0.53	0.68	0.59	0.71	0.66	0.72
North Carolina	133	0.65	0.61	0.53	0.65	0.57	0.68
North Dakota	120	0.42	0.45	0.24	0.42	0.54	0.49
Ohio	128	0.62	0.68	0.56	0.67	0.66	0.71
Oklahoma	129	0.47	0.47	0.43	0.49	0.48	0.56
Pennsylvania	127	0.59	0.66	0.54	0.72	0.63	0.69
Rhode Island	125	0.61	0.63	0.48	0.68	0.66	0.60
South Carolina	133	0.64	0.52	0.52	0.61	0.70	0.72
Tennessee	131	0.59	0.63	0.49	0.60	0.44	0.59
Texas	128	0.61	0.67	0.51	0.59	0.62	0.69
Utah	130	0.54	0.53	0.40	0.63	0.60	0.40
Virginia	128	0.60	0.55	0.55	0.64	0.62	0.69
Virgin Islands	98	0.52	0.64	0.49	0.76	0.70	0.64
West Virginia	127	0.59	0.60	0.45	0.57	0.57	0.44
Wisconsin	125	0.68	0.63	0.59	0.48	0.67	0.81
Wyoming	127	0.38	0.46	0.27	0.43	0.49	0.44

Table 9-12
Proportion of Proficiency Variance Accounted for by Grade 8 Conditioning Models

State	Number of Principal Components	Numbers and Operations	Measurement	Geometry	Data Analysis, Statistics, and Probability	Algebra and Functions	Estimation
Alabama	149	0.64	0.67	0.56	0.67	0.81	0.63
Arizona	148	0.62	0.64	0.56	0.64	0.75	0.59
Arkansas	148	0.65	0.61	0.58	0.66	0.78	0.63
California	150	0.68	0.66	0.63	0.70	0.81	0.67
Colorado	149	0.64	0.64	0.56	0.64	0.79	0.62
Connecticut	150	0.72	0.72	0.65	0.75	0.82	0.70
Delaware	129	0.70	0.71	0.63	0.73	0.84	0.73
District of Columbia	147	0.57	0.57	0.59	0.66	0.83	0.63
Florida	153	0.67	0.66	0.61	0.70	0.81	0.65
Georgia	152	0.64	0.65	0.62	0.66	0.80	0.64
Guam	106	0.77	0.71	0.59	0.77	0.85	0.70
Hawaii	140	0.59	0.59	0.64	0.70	0.84	0.69
Idaho	139	0.53	0.61	0.39	0.47	0.70	0.45
Indiana	150	0.65	0.64	0.64	0.70	0.81	0.66
Iowa	142	0.55	0.58	0.53	0.60	0.74	0.59
Kentucky	145	0.73	0.62	0.57	0.69	0.82	0.65
Louisiana	149	0.60	0.57	0.56	0.64	0.76	0.54
Maine	138	0.49	0.59	0.44	0.55	0.74	0.60
Maryland	145	0.76	0.74	0.74	0.78	0.83	0.72
Massachusetts	143	0.63	0.70	0.58	0.71	0.78	0.62
Michigan	149	0.64	0.63	0.61	0.63	0.80	0.65
Minnesota	141	0.57	0.56	0.64	0.63	0.79	0.59
Mississippi	151	0.62	0.61	0.53	0.78	0.83	0.65
Missouri	150	0.67	0.57	0.52	0.64	0.75	0.60
Nebraska	137	0.58	0.61	0.56	0.59	0.77	0.66
New Hampshire	137	0.65	0.59	0.52	0.60	0.74	0.56
New Jersey	145	0.68	0.72	0.70	0.72	0.82	0.69
New Mexico	145	0.66	0.65	0.52	0.70	0.77	0.59
New York	144	0.73	0.76	0.70	0.79	0.86	0.79
North Carolina	154	0.66	0.60	0.54	0.67	0.81	0.67
North Dakota	130	0.42	0.59	0.49	0.48	0.65	0.43
Ohio	145	0.73	0.66	0.64	0.76	0.82	0.71
Oklahoma	147	0.62	0.65	0.60	0.66	0.78	0.62
Pennsylvania	151	0.61	0.63	0.66	0.67	0.79	0.65
Rhode Island	134	0.69	0.68	0.56	0.72	0.79	0.66
South Carolina	151	0.70	0.72	0.68	0.73	0.82	0.68
Tennessee	149	0.64	0.64	0.56	0.69	0.78	0.65
Texas	152	0.68	0.70	0.62	0.68	0.78	0.59
Utah	148	0.56	0.60	0.47	0.56	0.77	0.60
Virginia	150	0.65	0.62	0.67	0.69	0.81	0.65
Virgin Islands	106	0.64	0.56	0.52	0.61	0.82	0.63
West Virginia	147	0.62	0.62	0.54	0.59	0.76	0.55
Wisconsin	142	0.55	0.58	0.60	0.64	0.78	0.62
Wyoming	134	0.62	0.56	0.49	0.62	0.75	0.56

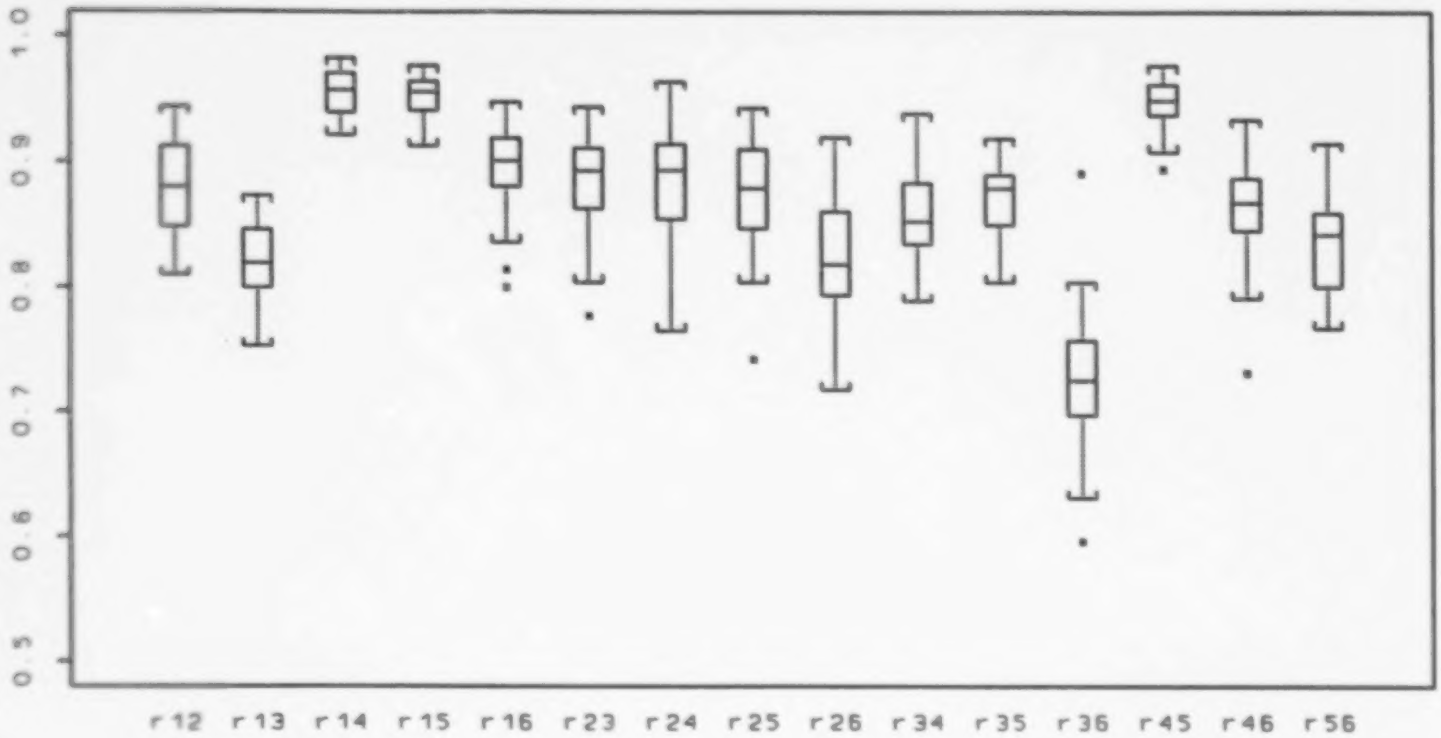
Figure 9-12
Boxplots of Estimated Scale Correlations*, Grade 4



---

* Each boxplot refers to the distribution (across participants) of estimated correlations between two of the six mathematics scales. The scales are identified by pairs of numbers across the bottom of the figure: 1 = Numbers and Operations; 2 = Measurement; 3 = Geometry; 4 = Data Analysis, Statistics, and Probability; 5 = Algebra and Functions; 6 = Estimation. For example, the boxplot identified as "r12" refers to the estimated correlations between the Numbers and Operations scale and the Measurement scale.

Figure 9-13
Boxplots of Estimated Scale Correlations*, Grade 8



---

* Each boxplot refers to the distribution (across participants) of estimated correlations between two of the six mathematics scales. The scales are identified by pairs of numbers across the bottom of the figure: 1 = Numbers and Operations; 2 = Measurement; 3 = Geometry; 4 = Data Analysis, Statistics, and Probability; 5 = Algebra and Functions; 6 = Estimation. For example, the boxplot identified as "r12" refers to the estimated correlations between the Numbers and Operations scale and the Measurement scale.

grades. At the fourth grade, and to a somewhat lesser extent at the eighth grade, the estimated correlations between Geometry and the remaining scales are noticeably lower than the correlations among the remaining scales and rarely exceed .9. In addition, there appears to be somewhat greater variability across jurisdictions in the correlations involving the Estimation scales, again with the effect being more clearly pronounced at the fourth grade. Furthermore, the correlation between the Geometry and Estimation scales is almost always the lowest among the set of correlations.

As discussed in Chapter 8, NAEP scales are viewed as summaries of consistencies and regularities that are present in item-level data. Such summaries should agree with other reasonable summaries of the item-level data. In order to evaluate the reasonableness of the scaling and estimation results, a variety of analyses were conducted to compare state-level and subgroup level performance in terms of the content area scaled scores and in terms of the average item score for the set of items in a content area. High agreement was found in all of these analyses. One set of such analyses is presented in Figures 9-14 and 9-15. The figures contain scatterplots of the state item score mean versus the state scale score means, for each of the five mathematics content areas. As is evident from the figures, there is an extremely strong relationship between the estimates of state-level performance in the scale-score and item-score metrics for all six content areas.

9.6 LINKING STATE AND NATIONAL SCALES

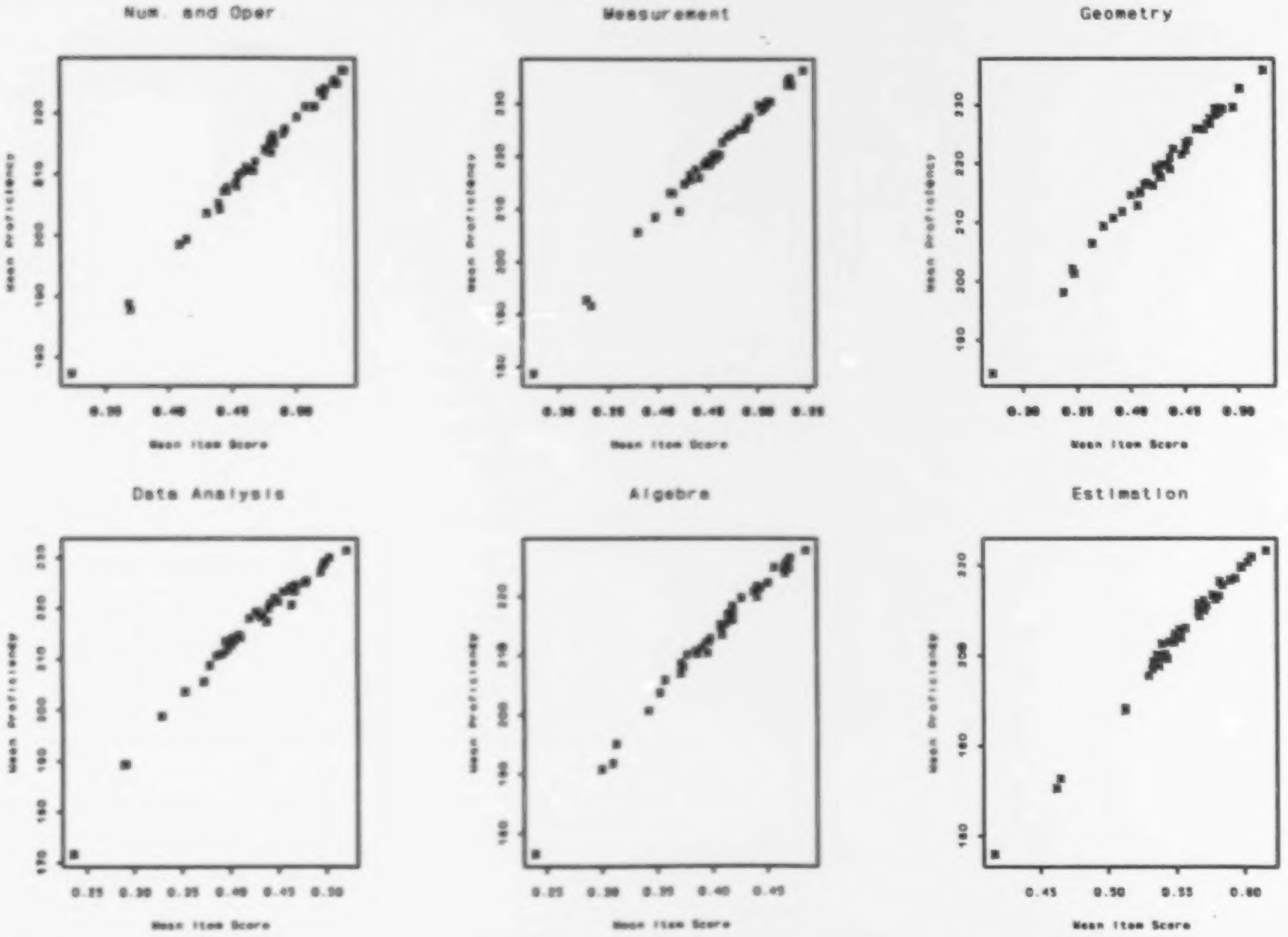
A major purpose of the Trial State Assessment Program was to allow each participating jurisdiction to compare its 1992 results at each grade level with the nation as a whole and with the region of the country in which that jurisdiction is located. Because 1992 was the second round of the Trial State Assessment, an additional goal was to provide an opportunity to compare 1992 results to those obtained in 1990 for those jurisdictions participating in both assessments.

For meaningful comparisons to be made between each of the Trial State Assessment participants and the relevant national samples, results from these two assessments had to be expressed in terms of a similar system of scale units. In addition, to allow for valid comparisons between grades, the systems of scale units for the fourth- and eighth-grade scales needed to be aligned and properly calibrated. Furthermore, the scales needed to be comparable to those used in 1990 to allow for meaningful assessment of changes in proficiency levels for jurisdictions participating in both assessments.

The fourth-grade and eighth-grade item pools did share a set of common items. However, as described in the previous section, separate scales were produced for the fourth and eighth grades in independent BILOG/PARSCALE calibrations. The units and origin of these scales were set by standardizing the within-grade proficiency distributions for their respective calibration samples to have a mean of zero and standard deviation of one. Thus, without further adjustment, the corresponding grade 4 and grade 8 scales were not expressed in similar systems of units. Some form of scale linking or calibration was required. In addition, although the fourth and eighth graders in the 1992 Trial State Assessment were administered the same test booklets as the fourth and eighth graders in the national assessment, separate state and national scalings were carried out (for reasons explained in Mazzeo, 1991, and Yamamoto and

Figure 9-14

Plot of Mean Proficiency Versus Mean Item Score, Grade 4



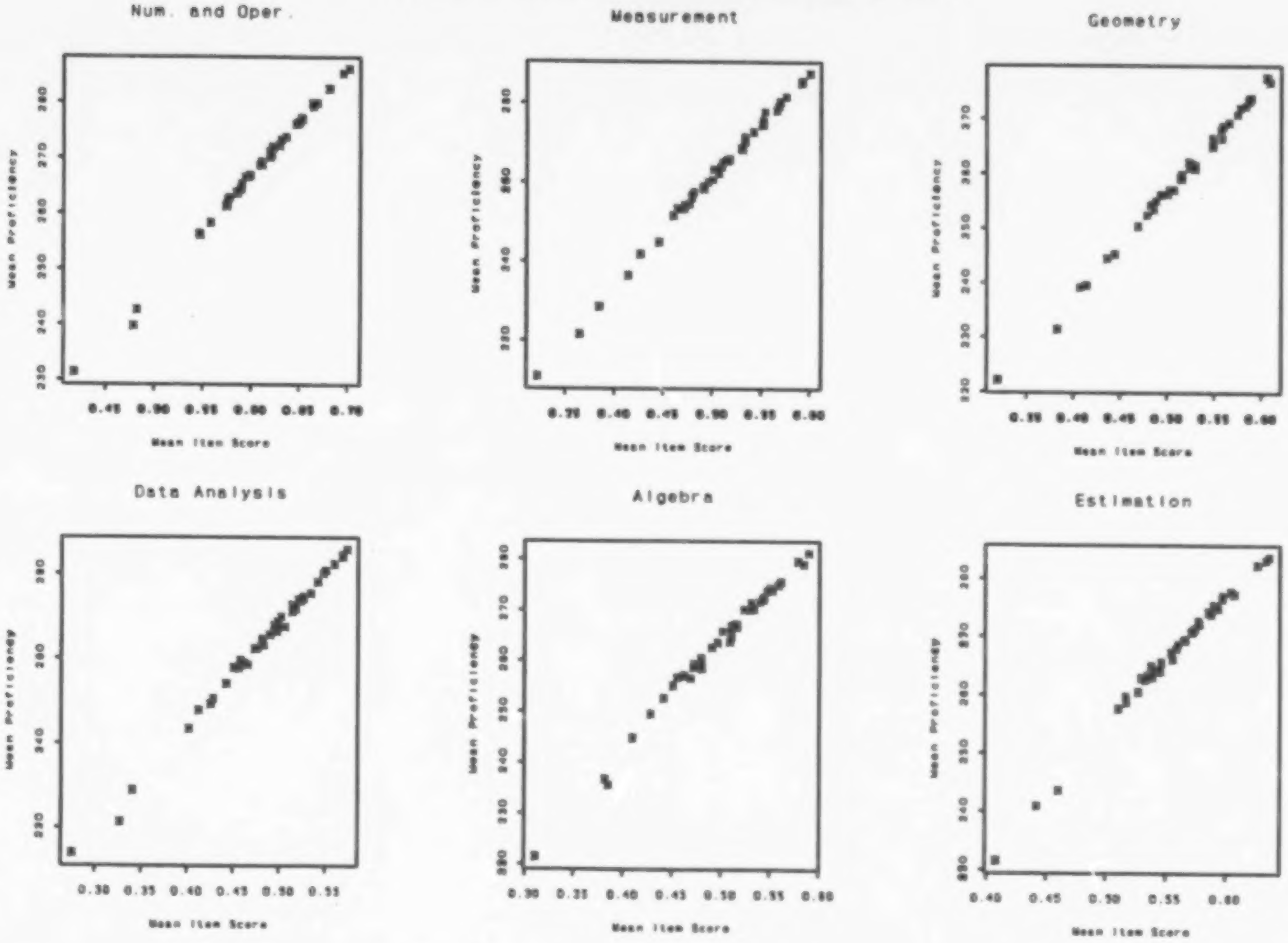
265

266

BEST COPY AVAILABLE

Figure 9-15

Plot of Mean Proficiency Versus Mean Item Score, Grade 8



Mazzeo, 1992). Again, to ensure a similar scale unit system for the state and national metrics, the metrics had to be linked. Plans for the scaling of the 1992 national assessment included procedures linking the 1992 scales to their 1990 counterparts. These procedures are described in the forthcoming technical report of the 1992 national assessment. Since the 1990 Trial State Assessment scales had already been linked to the 1990 national scales, linking the 1992 Trial State Assessment scales to their 1992 national counterparts indirectly linked the 1992 Trial State Assessment scales to the 1990 Trial State Assessment scales.

The purpose of this section is to describe the procedures used to align the 1992 Trial State scales with their 1992 national counterparts. The procedures that were used represent an extension of the common population equating procedures employed to link the 1990 national and state scales (Mazzeo, 1991; Yamamoto & Mazzeo, 1992).

Using the sampling weights provided by Westat, the combined sample of students from all participating jurisdictions was used to estimate the distribution of proficiencies for the population of students enrolled in public schools in the participating states and the District of Columbia[6]. Separate estimates were obtained for grades 4 and 8, with total sample sizes of 108,154 and 105,275, respectively. Data were also used from a subsample of the national assessment at grades 4 and 8, each consisting of grade-eligible public-school students from any of the 44 jurisdictions who participated in the 1992 Trial State Assessment, along with appropriate weights provided by Westat, to obtain estimates of the distribution of proficiency for the same target population. Again, separate estimates were produced for fourth grade (based on 5,198 students) and eighth grade (5,605 students).

Thus, for each of the 12 scales, two sets of proficiency distributions were obtained. One set, based on the sample of combined data from the Trial State Assessment (referred to as the Trial State Assessment Aggregate Sample) and using item parameter estimates and conditioning results from that assessment, was in the metric of the 1992 Trial State Assessment. The other, based on the sample from the 1992 national assessment (referred to as the State Aggregate Comparison, or SAC, sample) and obtained using item parameters and conditioning results from that assessment, was in the reporting metric of the 1992 national assessment. The latter metric had already been linked to the 1990 national reporting metric using the procedures described in the forthcoming technical report of the 1992 national assessment. The 12 Trial State Assessment and national scales were made comparable by constraining the mean and standard deviation of the two sets of estimates to be equal.

More specifically, the following steps were followed to linearly link the scales of the two assessments:

1) For each scale at each grade, estimates of the proficiency distribution for the Trial State Assessment Aggregate Sample was obtained using the full set of plausible values generated by the MGROUP program. The weights used were the final sampling weights provided by Westat, not the rescaled versions discussed

---

[6] Students from Guam and the Virgin Islands were excluded from the definition of this target population; hence, data from students from these jurisdictions were not included in the combined Trial State Assessment samples at grade 4 or grade 8.

in section 9.3. For each grade and each scale, the arithmetic mean of the five sets of plausible values was taken as the overall estimated mean and the geometric mean of the standard deviations of the five sets of plausible values was taken as the overall estimated standard deviation.

2) For each scale at each grade, the estimated proficiency distribution of the SAC sample was obtained, again using the full set of plausible values generated by the MGROUP program. The weights used were specially provided by Westat to allow for the estimation of proficiency for the same target population of students estimated by the state data. The means and standard deviations of the distributions (in the 1992 national reporting metric) for each scale at each grade were obtained for this sample in the same manner as described in step 1.

3) For each scale at each grade, a set of linear transformation coefficients were obtained to link the state scale to the corresponding national scale. The linking was of the form

$$Y^{*} = k_1 + k_2 Y$$

where

Y = a scale level in terms of the system of units of the provisional BILOG/PARSCALE scale of the Trial State Assessment scaling

Y^{*} = a scale level in terms of the system of units comparable to those used for reporting the 1992 national mathematics results

k_2 = $[\text{Standard-Deviation}_{SAC}]/[\text{Standard-Deviation}_{TSA}]$

k_1 = $\text{Mean}_{SAC} - k_2[\text{Mean}_{TSA}]$

The final conversion parameters for transforming plausible values from the provisional BILOG/PARSCALE scales to the final Trial State Assessment reporting scales are given in Table 9-13. All Trial State Assessment results are reported in terms of the Y^{*} metric.

It is important to re-emphasize two features of the linking procedures just described. First, the 1992 national scales had already been linked to their 1990 counterparts. Hence, the linking just described places the 1992 state scales on a metric comparable to that used for the 1990 national scales. Since the 1990 state metric was also made comparable to those same national scales, the 1992 and 1990 state results are in comparable metrics. Second, the 1990 national scales for each content area and for estimation were across-grade scales spanning grades 4, 8, and 12. Each had been produced by concurrently scaling the items from all three grade levels in a single BILOG calibration (see Yamamoto & Jenkins, 1992). For each content area, and for estimation, the grade 4 and grade 8 1992 state scales have been calibrated to the same 1990 across-grade scale. Hence, the grade 4 and grade 8 Trial State Assessment results are also on comparable scales.

Table 9-13

Transformation Constants for the Grade 4 and Grade 8 Scales

Scale	Grade 4		Grade 8	
	k_1	k_2	k_1	k_2
Numbers and Operations	215.53	34.78	268.25	34.95
Measurement	220.60	33.95	262.30	43.94
Geometry	219.95	29.16	260.16	34.20
Data Analysis, Statistics, and Probability	217.87	30.96	264.48	40.32
Algebra and Functions	217.00	29.66	263.61	36.61
Estimation	205.31	35.79	267.11	28.55

As evident from the discussion above, a linear method was used to link the scales from the Trial State and national assessments. While these linear methods ensure equality of means and standard deviations for the Trial State Assessment aggregate (after transformation) and the SAC samples, they do not guarantee the shapes of the estimated proficiency distributions for the two samples to be the same. As these two samples are both from a common target population, estimates of the proficiency distribution of that target population based on each of the samples should be quite similar in shape in order to justify strong claims of comparability for the Trial State and national scales. Substantial differences in the shapes of the two estimated distributions would result in differing estimates of the percentages of students above achievement levels or of percentile locations depending on whether Trial State or national scales were used—a clearly unacceptable result given claims about comparability of scales. In the face of such results, nonlinear linking methods would be required.

Analyses were carried out (one set of analyses for grade 4 and one set for grade 8) to verify the degree to which the linear linking process described above produced comparable scales for Trial State and national results. Comparisons were made between two estimated proficiency distributions, one based on the Trial State Assessment aggregate and one based on the SAC sample, for each of the six mathematics scales. The comparisons were carried out using slightly modified versions of what Wainer (1974) refers to as suspended rootograms. The final reporting scales for the Trial State and national assessments were each divided into 10-point intervals. Two sets of estimates of the percentage of students in each interval were obtained, one based on the Trial State Assessment aggregate sample and one based on the SAC sample. Following Tukey (1971), the square root of these estimated percentages were compared[7].

The comparisons are shown in Figures 9-16 through 9-21. The heights of each of the unshaded bars correspond to the square root of the percentage of students from the Trial State Assessment aggregate sample in each 10-point interval on the final reporting scale. The shaded bars show the differences in root percents between the Trial State Assessment and SAC estimates[8]. Positive differences indicate intervals in which the estimated percentages from the State Aggregate Comparison sample are lower than those obtained from the Trial State Assessment aggregate. Conversely, negative differences indicate intervals in which the estimated percentages from the State Aggregate Comparison sample are higher. For all six scales at both grades, differences in root percents are quite small, suggesting that the shapes of the two estimated distributions are quite similar (i.e., unimodal with slight negative skewness). There is some evidence that the estimates produced using the Trial State Assessment data are slightly heavier in the extreme lower tails (below 100 for the grade 4 scales and below 150 for the grade

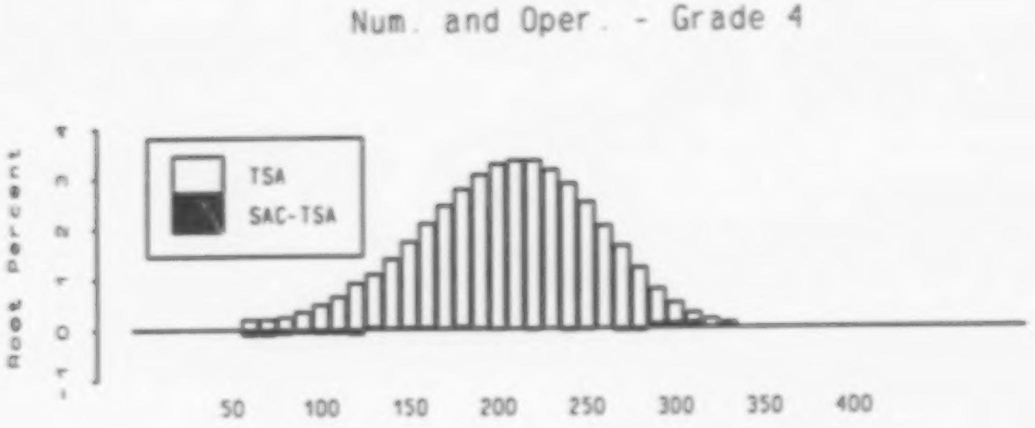
---

[7] The square root transformation allows for more effective comparisons for counts (or equivalently, percentages) when the expected number of counts in each interval is likely to vary greatly over the range of intervals, as is the case for the NAEP scales where the expected counts of individuals in intervals near the extremes of the scale (e.g., below 150 and above 350) are dramatically smaller than the counts obtained near the middle of the scale.

[8] Wainer (1974), among others, has suggested that looking at residuals around a fitted straight line makes judgments of differences somewhat easier to make. Hence, the *differences between the root percents*—rather than separate sets of root percents—from the SAC sample and the Trial State Assessment aggregate are plotted around the x-axis in Figures 9-16 through 9-21.

Figure 9-16

Rootogram Comparing Proficiency Distributions
for the Trial State Assessment Aggregate Sample
and the State Aggregate Comparison Sample from the National Assessment
for the Numbers and Operations Scale



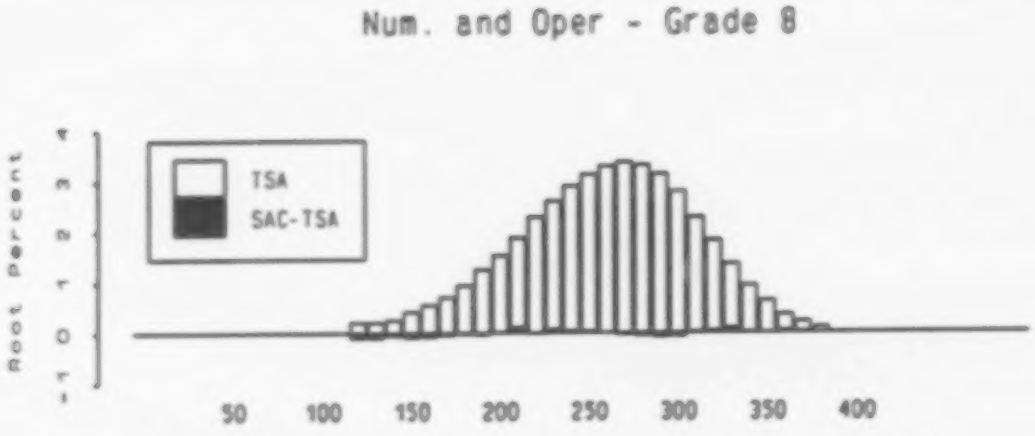
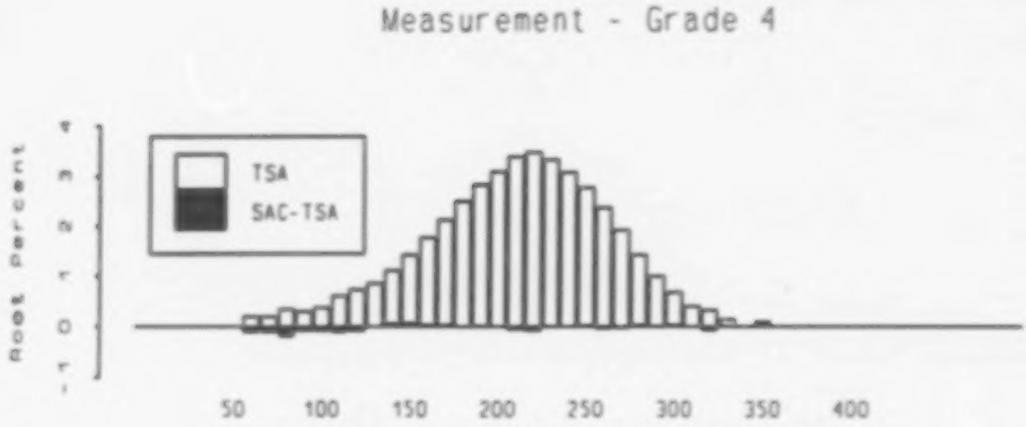


Figure 9-17

**Rootogram Comparing Proficiency Distributions
for the Trial State Assessment Aggregate Sample
and the State Aggregate Comparison Sample from the National Assessment
for the Measurement Scale**



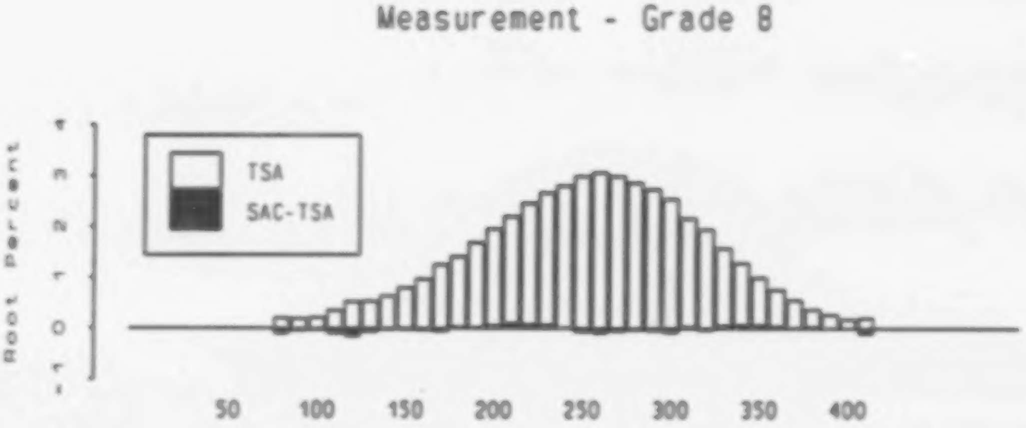
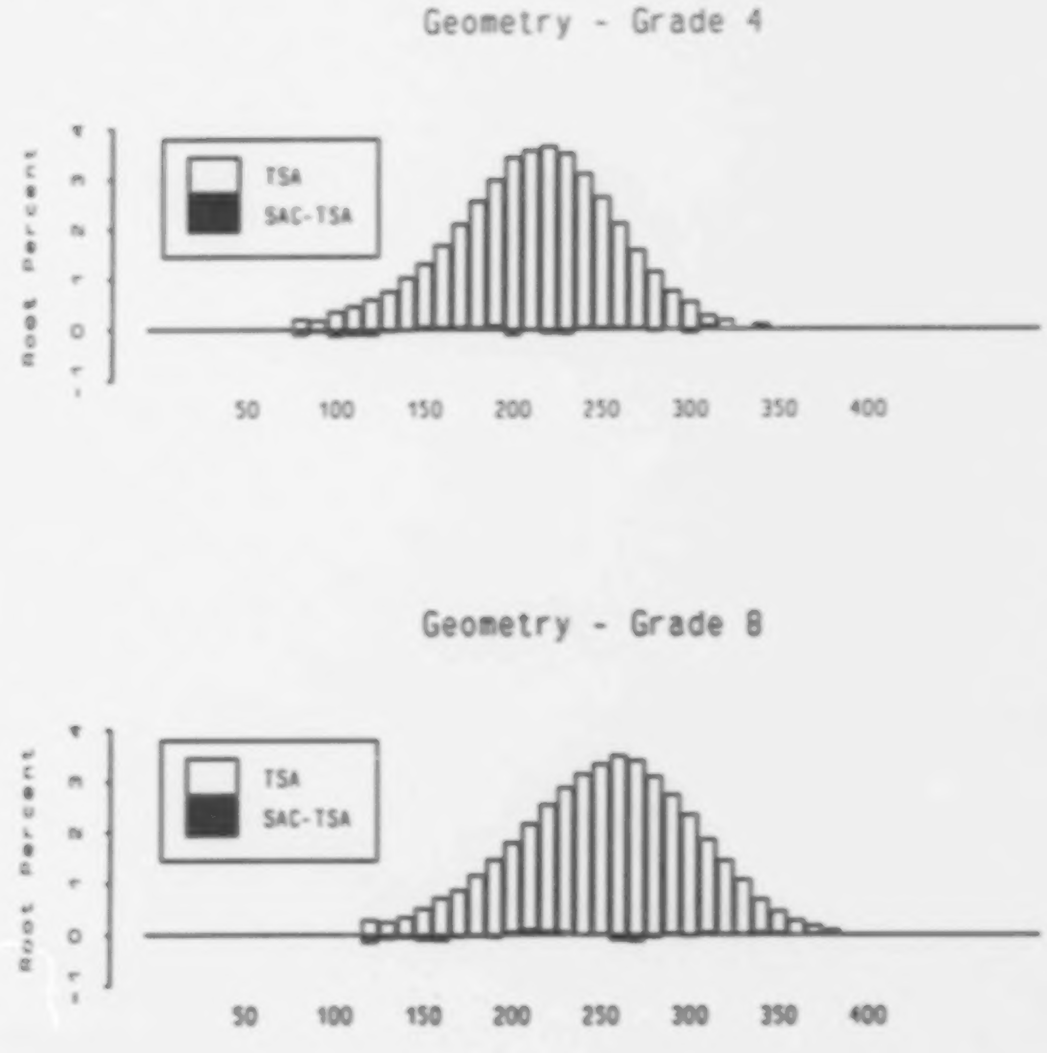


Figure 9-18

Rootogram Comparing Proficiency Distributions
for the Trial State Assessment Aggregate Sample
and the State Aggregate Comparison Sample from the National Assessment
for the Geometry Scale

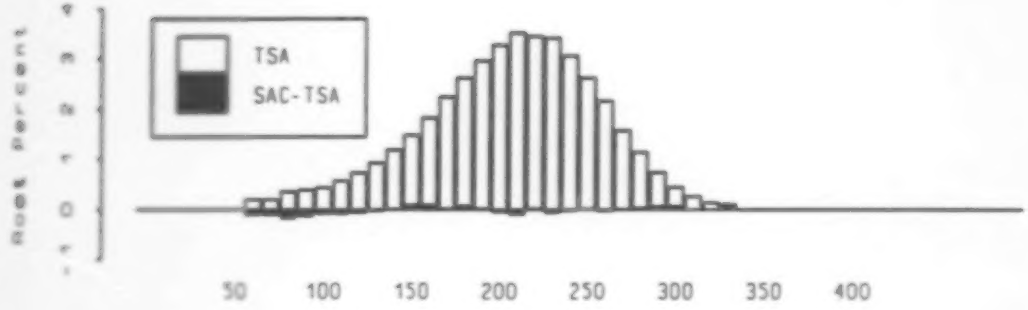

Geometry - Grade 4

Geometry - Grade 8

Figure 9-19

Rootogram Comparing Proficiency Distributions
for the Trial State Assessment Aggregate Sample
and the State Aggregate Comparison Sample from the National Assessment
for the Data Analysis, Statistics, and Probability Scale

Data Analysis - Grade 4



Data Analysis - Grade 8

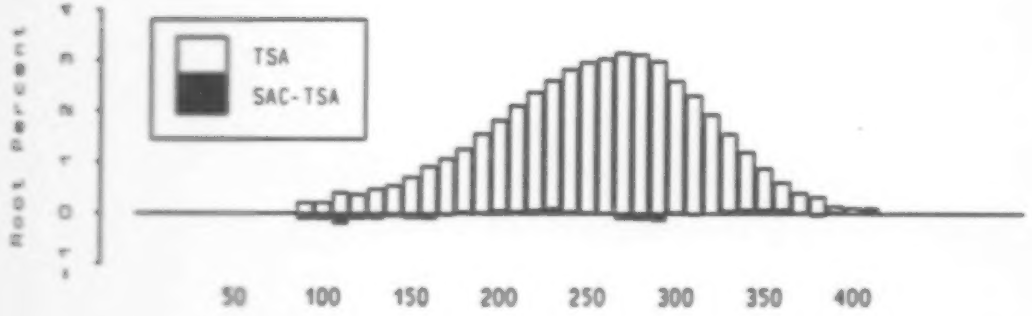
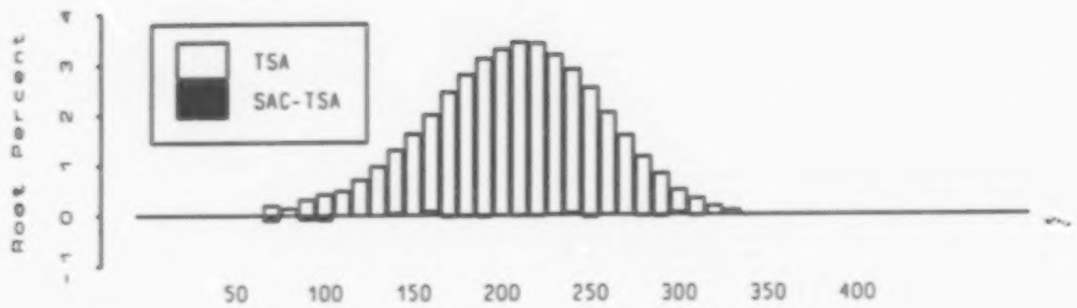


Figure 9-20

Rootogram Comparing Proficiency Distributions for the Trial State Assessment Aggregate Sample and the State Aggregate Comparison Sample from the National Assessment for the Algebra and Functions Scale

Algebra - Grade 4



Algebra - Grade 8

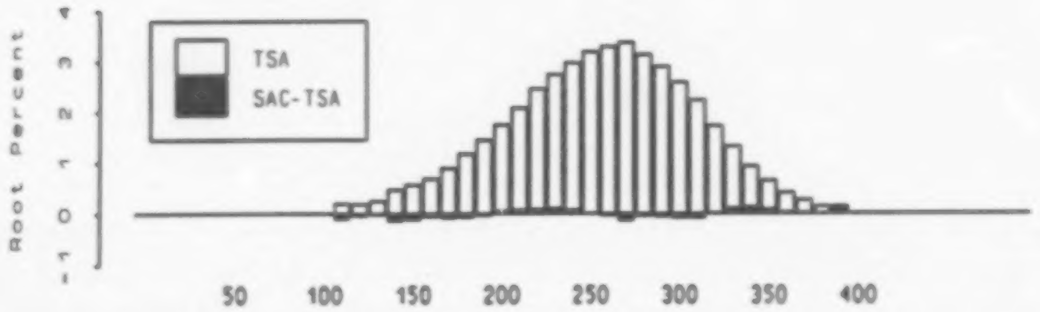
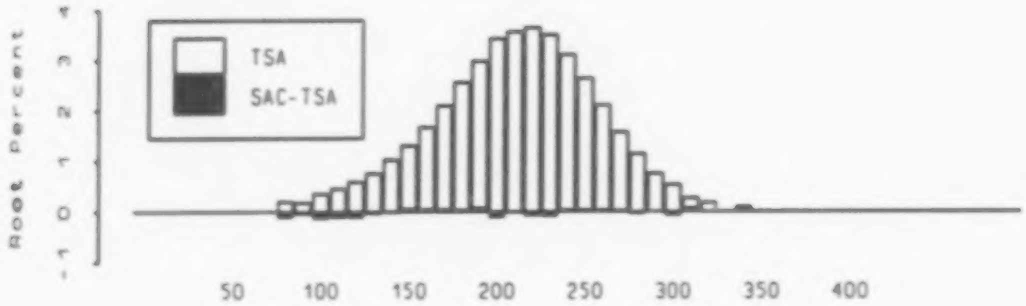


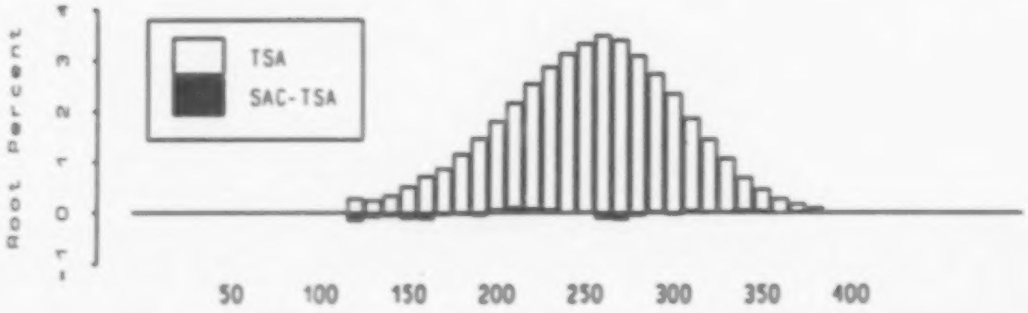
Figure 9-21

Rootogram Comparing Proficiency Distributions
for the Trial State Assessment Aggregate Sample
and the State Aggregate Comparison Sample from the National Assessment
for the Estimation Scale

Estimation - Grade 4



Estimation - Grade 8



8 scales). However, even these differences at the extremes are small in magnitude (.2 in the root percent metric, .04 in the percent metric) and have little impact on estimates of reported statistics such as percentages of students below the achievement levels.

9.7 PRODUCING A MATHEMATICS COMPOSITE SCALE

For the national assessment, composite scales were created for both fourth and eighth grade as overall measures of mathematics proficiency for students at that grade. The composite was a weighted average of plausible values on the five content area scales (Numbers and Operations; Measurement; Geometry; Data Analysis, Probability, and Statistics; and Algebra and Functions). The weights for the national content area scales were proportional to the relative importance assigned to each content area for each grade in the assessment specifications developed by the Mathematics Objectives Panel. Consequently, the weights for each of the content areas are similar to the actual proportion of items from that content area at each grade.

Trial State Assessment composite scales were developed using weights identical to those used to produce the composites for the 1992 national mathematics assessment. The weights are given in Table 9-14. In developing the Trial State Assessment composite for each grade, the weights were applied to the plausible values for each content area scale as expressed in terms of the final Trial State Assessment scales for each grade (i.e., after transformation from the provisional BILOG/PARSCALE scales.)

Figure 9-22 provides rootograms comparing the estimated proficiency distributions based on the Trial State Assessment and SAC samples for the grade 4 and grade 8 composites. Consistent with the results presented separately by scale, there is some evidence that the estimates produced using the Trial State Assessment data are slightly heavier in the extreme lower tails than the corresponding estimate based on the SAC data. However, again these differences in root relative percents are small in magnitude.

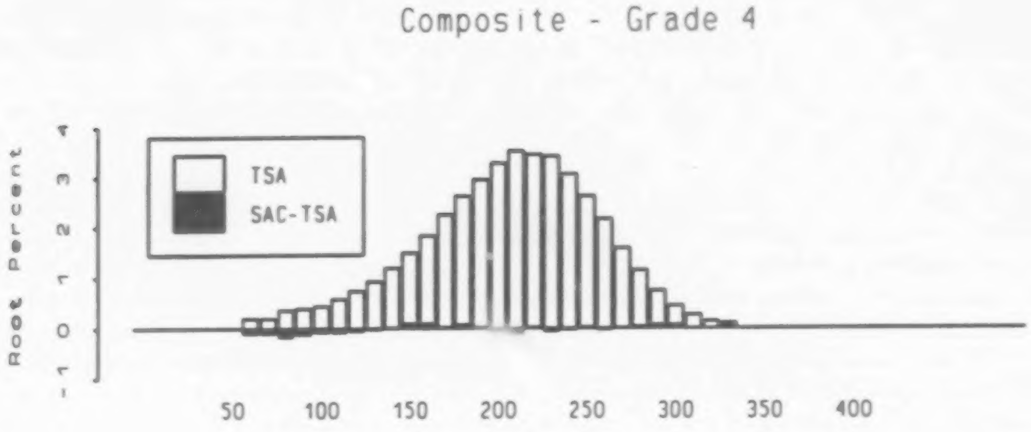
Table 9-14

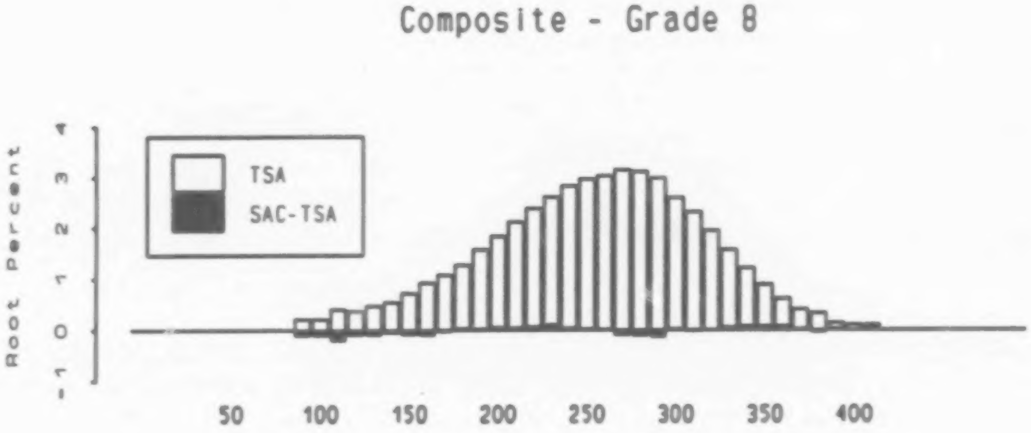
Weights Used for Each Scale to Form Grade 4 and Grade 8 Composites

Scale	Grade 4	Grade 8
Numbers and Operations	.45	.30
Measurement	.20	.15
Geometry	.15	.20
Data Analysis, Statistics, and Probability	.10	.15
Algebra and Functions	.10	.20

Figure 9-22

Rootogram Comparing Proficiency Distributions for the Trial State Assessment Aggregate Sample and the State Aggregate Comparison Sample from the National Assessment for the Composite Scale





Chapter 10

CONVENTIONS USED IN REPORTING THE RESULTS OF THE 1992 TRIAL STATE ASSESSMENT IN MATHEMATICS

John Mazzeo

Educational Testing Service

10.1 OVERVIEW

Results for the 1992 Trial State Assessment in Mathematics were disseminated in four different reports: a *Mathematics Report* for each state, the *NAEP 1992 Mathematics Report Card for the Nation and the States*, the *Data Compendium from the NAEP 1992 Mathematics Assessment for the Nation and the States*, and a five-section almanac of data for each state.

The *Mathematics Report* is a computer-generated report that provides, for each state, mathematics results for their fourth- and eighth-grade students. While national and regional results[1] are included for comparison purposes, the major focus of each of these computer-generated reports is the results for that particular jurisdiction. Data about school and student participation rates are reported for each jurisdiction to provide information about the generalizability of the results. School participation rates are reported both in terms of the initially selected samples of schools and in terms of the finally achieved samples, including replacement schools. Several different student participation rates are reported, including the overall rate, the percentage of students excluded from the assessment, and the exclusion rates for Limited English Proficiency (LEP) students and for students with Individualized Education Plans (IEP). In addition to 1992 results, the state reports contain comparisons of 1992 eighth-grade results to the 1990 eighth-grade results for the states and territories that participated in both assessments. Trend results are also provided for the nation and for the relevant region associated with each participant[2].

The state report text and tables were produced by a computerized report generation system developed by ETS report writers, statisticians, data analysts, graphic designers, and editors. Detailed technical documentation about the NAEP computer-generated reporting system can be found in the technical documentation of the 1992 NAEP computerized report generation system. The reports contain state-level estimates of proficiency means and standard deviations, proportions of students at or above achievement levels defined by the National Assessment Governing Board (NAGB) (see Appendix G for details on the definition and development of these levels), proportions of students at or above the traditional NAEP anchor levels, and selected percentiles for the state as a whole and for subgroups defined by four key reporting variables (referred to here as primary reporting variables)—gender, race/ethnicity, level of parents' education, and type of community. In addition, proficiency means are also reported for a variety of other subpopulations (referred to as secondary reporting variables) defined by responses to items from the student, teacher, and school questionnaires and by school and community demographic variables provided by Westat[3]. Item-level results are also provided for the set of extended constructed-response items included in the 1992 mathematics assessment.

---

[1] The national and regional results included in the state reports and in the portions of the *Data Compendium from the NAEP 1992 Mathematics Assessment for the Nation and the States* that present state results are based on data from the 1992 national mathematics assessment and include fourth-grade and eighth-grade students enrolled in public schools.

[2] All 1990 proficiency results (state, regional, and national) reported in the 1992 reports differ slightly from those originally reported. Improved proficiency estimation procedures were used in 1992. In order to maintain as much comparability as possible for trend comparisons, the 1990 data were reanalyzed using the new estimation procedures and the revised results were used for comparison to 1992. Further details can be found in Appendix H.

A second report, the *NAEP 1992 Mathematics Report Card for the Nation and the States*, highlights key assessment results for the nation and summarizes results across the states and territories participating in the assessment. This report contains composite scale results (proficiency means, proportions at or above achievement levels, etc.) for the nation, each of the four regions of the country, and each jurisdiction participating in the Trial State Assessment, both overall and by the primary reporting variables. In addition, overall results are reported for each of the content area scales. Reported results include trend comparisons to 1990 for all grades in the national assessment and for grade 8 for those jurisdictions that participated in both the 1990 and 1992 Trial State Assessments. The *Report Card* also contains a number of specially developed graphical displays that summarize and compare results for the full set of Trial State Assessment participants.

The third report is entitled *Data Compendium from the NAEP 1992 Mathematics Assessment for the Nation and the States*. Like the *Report Card*, the *Compendium* reports results for the nation and for all of the states and territories participating in the Trial State Assessment. However, unlike the *Report Card*, the *Compendium* is primarily tabular in nature and contains little in the way of interpretive text. The *Compendium* contains most of the tables included in the *Report Card* plus additional tables that provide composite scale results for a large number of secondary reporting variables. The variables used to report each jurisdiction's results in the individual state reports are included in the *Compendium*, along with additional background variables derived from the student, teacher, and school questionnaires.

The fourth report is a five-section almanac. Three of the sections of the almanac (referred to as proficiency sections) present analyses based on responses to each of the questionnaires (student, mathematics teacher, and school) administered as part of the Trial State Assessment. For most background questions contained in these questionnaires, the proportion of students responding to each option and the mathematics composite proficiency mean for these students are reported with their jackknifed standard errors. The student proficiency section of the almanac also contains selected percentiles and the percentages of students at or above achievement levels and anchor points. Results are provided for the total group of students in each participating jurisdiction, as well as for groups defined by the primary reporting variables (gender, race/ethnicity, type of community, and level of parents' education). The fourth section of the almanac, the scale section, reports proficiency means and associated standard errors for the five mathematics content-area scales and the estimation scale. Results in this section are also reported for the total group in each state, as well as for select subgroups of interest. The final section of the almanac, the "p-value" section, provides the total-group proportion of correct responses to each cognitive item included in the assessment.

---

[3] Some of these variables were used by Westat in developing the sampling frame for the assessment and in drawing the sample of participating schools.

The production of the state reports, the *Report Card*, the *Data Compendium*, and the almanacs required a large number of decisions about a variety of data analysis and statistical issues. For example, a wide variety of demographic profiles and instructional practices exist across the states and territories that participated in the Trial State Assessment. Given the sample sizes obtained for each state, certain categories of the reporting variables contained limited numbers of examinees. A decision was needed as to what constituted a sufficient sample size to permit the reliable reporting of subgroup results, and which, if any, estimates were sufficiently unreliable to need to be identified (or flagged) as a caution to readers. As a second example, the state report contained computer-generated text that described the results for a particular state and compared total and subgroup performance within the state to that of the region and nation. A number of inferential rules, based on logical and statistical considerations, had to be developed to ensure that the computer-generated reports were coherent from a substantive standpoint and were based on statistical principals of significance testing. As a third example, the *Report Card* contained tables that statistically compared performance between 1992 and 1990 for each of the participating states and territories. Practical multiple comparison procedures were required to control for Type I errors without paying too large a penalty with respect to the power to detect real and substantive differences.

The purpose of this chapter is to document the major conventions and statistical procedures used in generating the state reports, the *Report Card*, the *Data Compendium*, and the almanacs. The principal focus of this chapter is on conventions used in the production of the computer-generated state reports. However, sections 10.2 to 10.4 contain material applicable to all four summary reports. Additional details about procedures relevant to the *Report Card* and *Data Compendium* can be found in the text and technical appendices of those reports.

10.2 MINIMUM SAMPLE SIZES FOR REPORTING SUBGROUP RESULTS

In all four reports, estimates of quantities such as composite and content area proficiency means, percentages of students at or above the achievement levels, and percentages of students indicating particular levels of background variables (as measured in the student, teacher, and school questionnaires) are reported for the total population of fourth- and eighth-grade students in each jurisdiction, as well as for certain key subgroups of interest. The subgroups were defined by four primary NAEP reporting variables. NAEP reports results for five racial/ethnic subgroups (White, Black, Hispanic, Asian American/Pacific Islander, and American Indian/Alaskan Native), four types of communities (advantaged urban, disadvantaged urban, extreme rural, and other non-extreme communities), and four levels of parents' education (did not finish high school, high school graduate, some college, college graduate). However, in some jurisdictions, and for some regions of the country, sample sizes were not large enough to permit accurate estimation of proficiency and/or background variable results for one or more of the categories of these variables.

For results to be reported for any subgroup, a minimum sample size of 62 was required. This number was arrived at by determining the sample size required to detect an effect size of 0.2 with a probability of .8 or greater[4]. The effect size of 0.2 pertains to the "true" difference in mean proficiency between the subgroup in question and the total fourth- or eighth-grade public-school population in the state, divided by the standard deviation of proficiency in the total population. An effect size of 0.2 was chosen following Cohen (1977), who classifies effect sizes less than this magnitude as "small." The same convention was used in reporting the 1990 Trial State Assessment results.

The summary reports also include large numbers of tables that provide estimates of the proportion of the students responding to each category of a secondary reporting variable, as well as the mean proficiency of the students within each category. In several instances, the number of students in a particular category of these background variables was also less than 62. The same minimum sample size restriction of 62 was applied to these subgroups as well.

10.3 ESTIMATES OF STANDARD ERRORS WITH LARGE MEAN SQUARED ERRORS

Standard errors of mean proficiencies, proportions, and percentiles play an important role in interpreting subgroup results and comparing the performances of two or more subgroups. The jackknife standard errors reported by NAEP are statistics whose quality depends on certain features of the sample from which the estimate is obtained. In certain cases, typically when the number of students upon which the standard error is based is small or when this group of students all come from a small number of participating schools, the mean squared error[5] associated with the estimated standard errors may be quite large. In the summary reports, estimated standard errors subject to large mean squared errors are followed by the symbol "!".

The magnitude of the mean squared error associated with an estimated standard error for the mean or proportion of a group depends on the coefficient of variation (*CV*) of the estimated size of the population group, denoted as N (Cochran, 1977, section 6.3). The coefficient of variation is estimated by:

$$CV(\hat{N}) = \frac{SE(\hat{N})}{\hat{N}}$$

where $\hat{N}$ is a point estimate of N and $SE(\hat{N})$ is the jackknife standard error of $\hat{N}$.

---

[4] A design effect of 2 was assumed for this purpose, implying a sample design-based variance twice that of simple random sampling. This is consistent with previous NAEP experience (Johnson & Rust, 1992).

[5] The mean squared error of the estimated standard error is defined as $\mathcal{E}\,[\hat{S} - \sigma]^2$, where $\hat{S}$ is the estimated standard error, σ is the "true" standard error, and $\mathcal{E}$ is the expectation operator.

Experience with previous NAEP assessments suggests that when this coefficient exceeds 0.2, the mean squared error of the estimated standard errors of means and proportions based on samples of this size may be quite large. Therefore, the standard errors of means and proportions for all subgroups for which the coefficient of variation of the population size exceeds 0.2 are followed by "!" in the tables of all summary reports. These standard errors, and any confidence intervals or significance tests involving these standard errors, should be interpreted with caution. (Further discussion of this issue can be found in Johnson & Rust, 1992.)

10.4 TREATMENT OF MISSING DATA FROM THE STUDENT, TEACHER, AND SCHOOL QUESTIONNAIRES

Responses to the student, teacher, and school questionnaires played a prominent role in all reports. Although the return rate on all three types of questionnaire was high at both the fourth and eighth grades[6], there were missing data from each type.

For the questionnaires, the reported estimated percentages of students in the various categories of background variables, and the estimates of the mean proficiency of such groups, were based on only those students for whom data on the background variable were available. In the terminology of Little and Rubin (1987), the analyses pertaining to a particular background variable presented in the state reports and the *Data Compendium* assume the data are missing completely at random (i.e., the mechanism generating the missing data is independent of both the response to the particular background variables and to proficiency).

The estimates of proportions and proficiencies based on "missing-completely-at-random" assumptions are subject to potential nonresponse bias if, as may be the case, the assumptions are not correct. The amount of missing data was small (usually, less than 2 percent) for most of the variables obtained from the student and school questionnaires. For analyses based on these variables, reported results are subject to little, if any, nonresponse bias. However, for particular background questions from the student and school questionnaires, the level of nonresponse in certain jurisdictions was somewhat higher. As a result, the potential for nonresponse biases in the results of analyses based on this latter set of questions is also somewhat greater. Background questions for which more than 10 percent of the returned questionnaires were missing are identified in background almanacs produced for each jurisdiction. Again, results for analyses involving these questions should be interpreted with some degree of caution.

In order to analyze the relationships between teachers' questionnaire responses and their students' achievement, each teacher's questionnaire had to be matched to all of the students who were taught mathematics by that teacher. Table 10-1 provides the percentages of fourth-grade students that were matched to teacher questionnaires for each of the 44 jurisdictions that participated in the Trial State Assessment. Table 10-2 contains similar information for the

---

[6] Information about survey participation rates (both school and student), as well as proportions of students excluded by each state from the assessment, are given in Appendix B. Adjustments intended to account for school and student nonresponse are described in Chapter 8.

Table 10-1
Weighted Percentage of Students Matched to Teacher Questionnaires for Grade 4

State	Questionnaire Match Rate		
	No Match	Part I Only	Parts I and II
Alabama	5.1	94.9	91.0
Arizona	2.0	98.0	92.5
Arkansas	3.1	96.9	94.3
California	2.7	97.3	90.6
Colorado	6.3	93.7	86.3
Connecticut	6.8	93.2	89.0
Delaware	2.5	97.5	94.3
District of Columbia	11.9	88.1	80.0
Florida	7.1	92.9	90.1
Georgia	4.8	95.2	88.2
Guam	2.4	97.6	97.5
Hawaii	3.4	96.6	92.5
Idaho	1.2	98.8	91.8
Indiana	3.7	96.3	89.3
Iowa	4.6	95.4	90.8
Kentucky	3.9	96.1	89.9
Louisiana	5.6	94.4	91.2
Maine	5.9	94.1	90.6
Maryland	3.4	96.6	90.2
Massachusetts	5.0	95.0	87.4
Michigan	5.1	94.9	89.9
Minnesota	7.9	92.1	80.4
Mississippi	4.9	95.1	88.6
Missouri	1.6	98.4	90.6
Nebraska	3.3	96.7	86.1
New Hampshire	2.6	97.4	94.9
New Jersey	1.4	98.6	92.5
New Mexico	6.8	93.2	87.3
New York	4.3	95.7	91.7
North Carolina	4.1	95.9	94.4
North Dakota	0.7	99.3	93.1
Ohio	5.0	95.0	89.6
Oklahoma	1.4	98.6	94.5
Pennsylvania	1.9	98.1	93.8
Rhode Island	3.7	96.3	92.1
South Carolina	1.0	99.0	97.3
Tennessee	1.2	98.8	93.4
Texas	6.8	93.2	84.5
Utah	1.3	98.7	95.2
Virginia	6.2	93.8	88.4
Virgin Islands	4.5	95.5	83.4
West Virginia	3.7	96.3	88.5
Wisconsin	2.6	97.4	90.2
Wyoming	4.8	95.2	91.4

BEST COPY AVAILABLE

Table 10-2
Weighted Percentage of Students Matched to Teacher Questionnaires for Grade 8

State	Questionnaire Match Rate		
	No Match	Part I Only	Parts I and II
Alabama	1.8	98.2	94.1
Arizona	1.6	98.4	91.1
Arkansas	2.2	97.8	95.0
California	2.9	97.1	92.5
Colorado	1.9	98.1	90.5
Connecticut	0.4	99.6	97.3
Delaware	0.3	99.7	97.4
District of Columbia	5.0	95.0	83.4
Florida	2.5	97.5	94.8
Georgia	1.8	98.2	94.1
Guam	7.5	92.5	90.8
Hawaii	1.7	98.3	89.8
Idaho	3.8	96.2	90.8
Indiana	1.9	98.1	92.4
Iowa	2.0	98.0	92.4
Kentucky	5.6	94.4	91.0
Louisiana	3.0	97.0	93.8
Maine	0.3	99.7	91.4
Maryland	4.2	95.8	90.9
Massachusetts	1.3	98.7	89.7
Michigan	3.1	96.9	93.2
Minnesota	5.9	94.1	83.8
Mississippi	1.4	98.6	95.1
Missouri	0.8	99.2	95.9
Nebraska	1.6	98.4	93.8
New Hampshire	2.1	97.9	92.9
New Jersey	0.3	99.7	96.2
New Mexico	2.5	97.5	92.9
New York	0.7	99.3	94.6
North Carolina	1.7	98.3	95.3
North Dakota	0.3	99.7	96.9
Ohio	3.1	96.9	89.6
Oklahoma	1.4	98.6	91.7
Pennsylvania	0.3	99.7	97.2
Rhode Island	3.4	96.6	82.8
South Carolina	2.1	97.9	94.6
Tennessee	1.6	98.4	94.7
Texas	1.1	98.9	94.5
Utah	5.2	94.8	90.2
Virginia	1.3	98.7	96.2
Virgin Islands	6.2	93.8	84.3
West Virginia	0.3	99.7	93.3
Wisconsin	6.1	93.9	89.8
Wyoming	3.0	97.0	88.8

BEST COPY AVAILABLE

eighth-grade samples. In both tables, three separate match rates are given. The first is the percentage of students that could not be matched to either part of the two-part teacher questionnaire. The second match rate is the percentage of students that could be matched to only the first part of the teacher questionnaire. The third is the percentage of students that could be matched to both the first and second parts of the teacher questionnaire. Note that these match rates do not reflect the additional missing data due to item-level nonresponse. The amount of additional item-level nonresponse in the returned teacher questionnaires can also be found in the almanacs produced for each jurisdiction.

10.5 STATISTICAL RULES USED FOR PRODUCING THE STATE REPORTS

As described earlier, the state reports contain state-level estimates of fourth- and eighth-grade mean proficiencies, proportions of students at or above selected scale points, and percentiles for the state as a whole and for the categories of a large number of reporting variables. Similar results are provided for the nation and, where sample sizes permitted, for the region to which each state belongs[7]. The state reports were computer-generated. The tables and figures, as well as the text of the report, were automatically tailored for each jurisdiction based on the pattern of results obtained. The purpose of this section is to describe some of the procedures and rules used to produce these individually tailored reports. A complete and detailed presentation is available in the forthcoming technical documentation of the 1992 NAEP computerized report generation system.

In the 1992 state reports, the results are presented principally through a sequence of tables containing estimated means, proportions, and percentiles, along with their standard errors, for 1992 and, where appropriate, for 1990. In addition to the tables of results, computer-generated interpretive text is also provided. In some cases, the computer-generated interpretive text is primarily descriptive in nature and reports the total group and subgroup proficiency means and proportions of interest. However, some of the interpretive text focuses on interesting and potentially important group differences in mathematics proficiency or on the percentages of students responding in particular ways to the background questions. Additional interpretive text compares state-level results with those of the nation, and discusses changes in results from 1990 to 1992. For example, one question of considerable interest to each jurisdiction is whether, on average, its students performed higher than, lower than, or about the same as students in the nation. Another question of interest is whether students from disadvantaged urban areas were less likely to be enrolled in an eighth-grade algebra course than were students from advantaged urban areas. Still another question of interest involves whether 1992 students evidenced higher levels of mathematics achievement than their 1990 counterparts. Additional interpretive text focuses on potentially interesting patterns of achievement across the five mathematics content areas or on the pattern of response to a particular background question in the state. For example, do more students report spending 30 minutes or 15 minutes on homework each day?

---

[7] Because United States territories are not classified into NAEP regions, no regional comparisons were provided for Guam and the Virgin Islands.

Rules were developed to produce the computer-generated text for questions involving the comparison of results for subgroups and interpretations of patterns of results. These rules were based on a variety of considerations, including a desire for 1) statistical rigor in the identification of important group differences and patterns of results, and 2) solutions that were within the limitations imposed by the availability of computational resources and the time frame for the production of the report. The following sections describe some of these procedures and rules.

10.5.1 Comparing Means and Proportions for Mutually Exclusive Groups of Students

Many of the group comparisons explicitly commented on in the state reports involved mutually exclusive sets of students. One common example of such a comparison is the contrast between the mean composite proficiency in a particular state and the mean composite proficiency in the nation. Other examples include comparisons within a jurisdiction of the average proficiency for male and female students; White and Hispanic students; students from advantaged urban schools and disadvantaged urban schools; and students who reported watching six or more hours of television each night and students who report watching less than one hour each night.

In the state reports, computer-generated text indicated that means or proportions from two groups were different only when the difference in the point estimates for the groups being compared was statistically significant at an approximate α level of .05. An approximate procedure was used for determining statistical significance that NAEP staff felt was reasonable from a statistical standpoint, as well as being computationally tractable. The procedure was as follows.

Let t_i be the statistic in question (i.e., a mean or proportion for group i) and let $SE(t_i)$ be the jackknife standard error of the statistic. The computer-generated text in the state report identified the means or proportions for groups i and j as being different if and only if:

$$\frac{|t_i - t_j|}{\sqrt{\hat{SE}^2(t_i) + \hat{SE}^2(t_j)}} \geq Z_{\frac{.05}{2c}}$$

where Z_α is the (1 - α) percentile of the standard normal distribution, and c is the number of contrasts being tested. In cases where group comparisons were treated as individual units (for example, comparing overall state results with overall national results or overall state results in 1992 with those of 1990, the value of c was taken as 1, and the test statistic was approximately equivalent to a standard two-tailed t-test for the difference between group means or proportions from large independent samples with the α level set at .05.

The procedures in this section assume that the data being compared are from independent samples. Because of the sampling design used for the Trial State Assessment, in which both schools and students within schools are randomly sampled, the data from mutually exclusive sets of students within a state may not be strictly independent. Therefore, the significance tests employed are, in many cases, only approximate. As described in the next section, another procedure, one that does not assume independence, could have been conducted.

However, that procedure is computationally burdensome and resources precluded its application for all the comparisons in the state reports. It was the judgment of NAEP staff that if the data were correlated across groups, in most cases the correlation was likely to be positive. Since, in such instances, significance tests based on assumptions of independent samples are conservative (because the estimated standard error of the difference based on independence assumptions is larger than the more complicated estimate based on correlated groups), the approximate procedure was used for most comparisons.

The procedures described above were also used for testing differences of both means *and* proportions. The approximation for the test for proportions works best when sample sizes are large, and the proportions being tested have magnitude close to .5. Statements about group differences should be interpreted with caution if at least one of the groups being compared is small in size and/or if somewhat extreme proportions are being compared.

10.5.2 Multiple Comparison Procedures

Frequently, groups (or families) of comparisons were made and were presented as a single set. The appropriate text, usually a set of sentences or a paragraph, was selected for inclusion in the report based on the pattern of results for the entire set of comparisons. For example, in Chapter 1 of the state report, state/territory results were compared to national results for each of the six content area scales. For families of contrasts like these, a Bonferroni procedure was used for determining the value of Z_c, where *c* was the number of contrasts in the set. In this example, *c* was taken to be 6, and each statistical test was consequently carried out at an α level of .05/6. As a second example, Chapter 2 of the state report contained a section that compared average proficiencies for a majority group (in the case of race/ethnicity, for example, usually White students) to those obtained by each minority group containing 62 or more students. Assuming three such minority groups, the text in the section was based on the results of three predefined statistical tests (i.e., a test comparing majority group performance to that of each of the three minority groups). Each statistical test was carried out at an α level of .05/3.

It should be noted that sets of statistical tests like those described in the paragraph above were carried out based on both fourth-grade and eighth-grade results. For the purposes of determining family sizes for Bonferroni adjustments, each grade was considered a separate set of contrasts. In the examples above, state-to-nation comparisons of proficiency means for each of the six scales were carried out at an α level of .05/6 for both fourth and eighth grades. Similarly, significance tests comparing a majority group to each of three minority groups were carried out at an α level of .05/3 at both grades.

10.5.3 Determining the Highest and Lowest Scoring Groups from a Set of Ranked Groups

Three analyses in the state report consisted of determining which of a set of several groups had the highest or lowest proficiency among the set. For example, one analysis compared the average proficiency of students who reported watching various amounts of television each day. There were five levels of television watching—one hour or less, two hours, three hours, four to five hours, and six hours or more. Based on their answers to this question in the student background questionnaire, students were classified into one of the five levels of television watching, and the mean composite proficiency was obtained for students at each level. The analysis focused on which, if any, of the groups had the highest and lowest mean composite proficiency.

The analysis was carried out using the statistics described in the previous section. The groups were ranked from highest to lowest in terms of their estimated mean proficiency. Then, three separate significance tests were carried out: 1) the highest group was compared to the lowest group; 2) the highest group was compared to the second highest group; and 3) the lowest group was compared to the second lowest group. The following conclusions were drawn:

- If all three comparisons were statistically significant, the performance of the highest ranking group was described as *highest* and the performance of the lowest ranking group was described as *lowest*.
- If only the first and second tests were significant, the highest ranking group was described as *highest*, but no comment was made about the lowest ranking group.
- Similarly, if only the first and third tests were significant, the lowest ranking group was described as *lowest*, but no comment was made about the highest ranking group.
- If only the first test was significant, the highest group was described as performing better than the lowest group, but no *highest* and *lowest* group were designated.

The Bonferroni adjustment factor was taken as the number of possible pairwise comparisons because of the ranking of groups prior to the carrying out of significance tests.

10.5.4 Comparing 1992 and 1990 Results in State Report Tables

Since its inception, one of NAEP's central purposes has been the monitoring of trends in achievement. The 1992 Trial State Assessment provided the first opportunity to report on short-term trends (from 1990 to 1992) in eighth-grade mathematics achievement and instructional practices on a state-by-state basis, as well as for the nation and the relevant region of the country. As a result, one of the prominent features of the 1992 state report was the inclusion of a large number of trend comparisons in both the text and tables of the reports for those jurisdictions that participated in both the 1990 and 1992 Trial State Assessments.

The samples for the 1990 and 1992 Trial State Assessments were drawn independently and consisted of mutually exclusive groups of students. Therefore, the selections of text describing comparisons of 1990 and 1992 results were based on the types of significance testing procedures described in section 10.5.1. In sections of the report where trend comparisons were carried out for a number of subgroups (e.g., where 1992 results were compared to 1990 results for each race/ethnicity group within the state, or for each of the content area scales), the significance testing procedures incorporated Bonferroni adjustments, like those described in section 10.5.2, which were based on the number of comparisons being made.

In addition, a large number of state report tables provided both 1990 and 1992 percentages of students and proficiency means for the subgroups of students defined by primary and secondary reporting variables. In most of these tables, three sets of trend results were reported, one set for the state/territory in question, one set for relevant region of the country, and one set for the nation. For each of these sets of results, symbols were included next to the 1992 results for each jurisdiction indicating which, if any, of the reported statistics represented a significant change from the 1990 results. A ">" sign was used to indicate 1992 results that were significantly higher than their corresponding 1990 levels. A "<" was used to indicate 1992 results that were significantly lower than their corresponding 1990 levels. No symbol appeared after results that did not differ significantly from their 1990 levels.

As was done for text selection, statistical tests were carried out using Bonferroni adjustments to significance levels when results for multiple groups were included in a table. For example, in a table containing 1990 and 1992 mean proficiencies for White, Black, and Hispanic students, statistical tests for differences were carried out at an α level of .05/3. It should be noted that national, regional, and state/territory comparisons were treated as separate families for the purposes of obtaining Bonferroni adjustments. Continuing with the race/ethnicity example, state/territory, national, and regional comparisons were treated as three separate families each consisting of three comparisons and each of the required statistical tests were carried out at an α level of .05/3.

10.5.5 Comparing Dependent Proportions

Certain analyses in the state report involved the comparison of dependent proportions. One example was the comparison of the proportion of students who reported that they spent 30 minutes a day on homework to the proportion of students who indicated that they spent 15 minutes a day on homework to determine which proportion was larger. For these types of analyses, NAEP staff determined that the dependencies in the data could not be ignored.

Unlike the case for analyses of the type described in section 10.5.1, the correlation between the proportion of students reporting 30 minutes of homework and the proportion reporting 15 minutes is likely to be negative. For a particular sample of students, it is likely that the higher the proportion of students reporting 30 minutes is, the lower the proportion of students reporting 15 minutes will be. A negative dependence will result in underestimates of the standard error if the estimation is based on independence assumptions (as is the case for the procedures described in the previous section). Such underestimation can result in too many "nonsignificant" differences being identified as significant.

The procedures of section 10.5.1 were modified for the state report analyses that involved comparisons of dependent proportions. The modification involved using a jackknife method for obtaining the standard error of the difference in dependent proportions. The standard error of the difference in proportions was obtained by first obtaining a separate estimate of the difference in question for each jackknife replicate, using the first plausible value only, then taking the standard deviation of the set of replicate estimates as the estimate. The procedures used for dependent proportions differed from the procedures of section 10.5.1 only with respect to estimating the standard error of the difference; all other aspects of the procedures were identical.

10.5.6 Statistical Significance and Estimated Effect Sizes

Whenever single comparisons were made between groups, an attempt was made to distinguish between group differences that were statistically significant but rather small in a practical sense and differences that were both statistically and practically significant. In order to make such distinctions, a procedure based on estimated effect sizes was used. The estimated effect size for comparing means from two groups was defined as:

$$\textit{estimated effect size} = \frac{|\hat{\mu}_i - \hat{\mu}_j|}{\sqrt{\dfrac{S_i^2 + S_j^2}{2}}}$$

where $\hat{\mu}_i$ refers to the estimated mean for group *i*, and S_i refers to the estimated standard deviation within group *i*. The within-group estimated standard deviations were taken to be the standard deviation of the set of five plausible values for the students in subgroup *i* and were calculated using the Westat sampling weights.

The estimated effect size for comparing proportions was defined as $|f_i - f_j|$, where $f_i = 2 \arcsin\sqrt{p_i}$ and p_i is the estimated proportion in group *i*.

For both means and proportions, no qualifying language was used in describing significant group differences when the estimated effect size exceeded .1. However, when a significant difference was found but the estimated effect size was less than .1, the qualifier *somewhat* was used. For example, if the mean proficiency for females was significantly higher than that for males but the estimated effect size of the difference was less than .1, females were described as performing *somewhat higher* than males.

The principal audience for the state reports was taken to consist of curriculum- and policy-oriented education specialists. Although it was assumed that such an audience would have some degree of familiarity with statistics, an attempt was made to keep the amount of statistical jargon to a minimum. This caused a certain degree of difficulty for group comparisons in which no statistically significant difference was obtained. In such cases, the rigorous statistical interpretation is not that the groups are the same (one does not prove the null hypothesis), but that the data are not sufficiently strong to justify concluding that a difference exists. In order to minimize the use of phrases such as "no statistically significant difference," the performance levels of the groups being compared were sometimes described as being "about the same". Readers were cautioned in the introduction to the state reports to interpret such statements to mean "no statistically significant difference."

The reliance on significance tests for commenting on differences while adopting a convention of describing null results as "about the same" resulted in situations that might appear somewhat anomalous to a reader of the report. Due to variations in subgroup sample sizes and standard errors, group differences between point estimates of one quantity (like a subgroup mean or proportions) could be large in an absolute sense but not statistically different (and hence described as "about the same") while a considerably smaller difference between another pair of groups was described as indicating different levels of performance. An attempt was made to minimize potential confusion by footnoting large but nonsignificant differences. If the difference in proficiency means between two groups was greater than 7 points, a footnote appeared on the page on which the comparison was described. The footnote read, "Recall that 'about the same' means that the difference between groups, although it may appear large, is not statistically significant."

10.5.7 Descriptions of the Magnitude of Percentage

Percentages reported in the text of the state reports are sometimes described using quantitative words or phrases. For example, the number of students being taught by teachers with master's degrees in mathematics might be described as "relatively few" or "almost all," depending on the size of the percentage in question. Any convention for choosing descriptive terms for the magnitude of percentages is to some degree arbitrary. The rules used to select the descriptive phrases in the report are given in Table 10-3.

Table 10-3
Rules for Selecting Descriptions of Percentages

Percentage	Descriptive Text Used in Report
$p = 0$	None
$0 < p \leq 10$	Relatively few
$10 < p \leq 20$	Some
$20 < p \leq 30$	About one-quarter
$30 < p \leq 44$	Less than half
$44 < p \leq 55$	About half
$55 < p \leq 69$	More than half
$69 < p \leq 79$	About three-quarters
$79 < p \leq 89$	Many
$89 < p < 100$	Almost all
$p = 100$	All

10.6 Comparisons of 1992 and 1990 Eighth-grade Results in the *Mathematics Report Card* and the *Data Compendium*

Both the *Mathematics Report Card* and the *Data Compendium* contain a large number of tables that compare eighth-grade results for 1992 with those obtained in 1990 for the nation as a whole, for each of the four regions of the country, and for each of the 37 jurisdictions that participated in both the 1990 and 1992 Trial State Assessments. The national and regional results are based on the 1990 and 1992 national NAEP samples, excluding private school students (both years) and students tested during the spring administration cycle (1990 only). The results for the states and territories are based on the 1992 and 1990 Trial State Assessment samples. Each jurisdiction's overall results are compared, as well as the results for both primary and secondary NAEP reporting subgroups. The following statistics are compared:

- the proportions of examinees in the various primary and secondary reporting subgroups;
- average proficiencies, overall and for the primary reporting subgroups, on the composite scale, estimation scale, and the five content area scales;
- proportions of students at or above the achievement levels, overall and within the primary reporting subgroups, on the composite and estimation scales;
- selected percentiles (5th, 10th, 25th, 50th, 75th, 90th, and 95th) overall, for the NAEP composite scale, estimation scale, and each of the five content area scales; and
- proportions of students at or above the achievement levels, overall and within the primary reporting subgroups, on the composite scale.

A number of different types of tables are included in the *Mathematics Report Card* and the *Data Compendium*. For example, one type of table shows the average composite proficiency and the percentage of students at or above each of the achievement levels. A second type of table shows the percentage of students at or above achievement levels on the composite scales for each of the primary reporting subgroups. A third type of table shows average mathematics proficiency and seven percentile locations for each of the five content area scales. A fourth type of table shows average composite proficiencies for a particular set of primary or secondary reporting subgroups.

Because of the large volume of tables in the *Mathematics Report Card* and the *Data Compendium*, most were computer-generated. To help readers focus on important outcomes, each of the tables containing eighth-grade results for both 1992 and 1990 are annotated with symbols indicating which 1992-to-1990 state (or territory) comparisons represent statistically significant changes[8]. The annotations to these tables were made automatically by the computer programs that produced them and were based on tests of statistical significance and Bonferroni adjustments like those described in sections 10.5.1 and 10.5.2. This section describes the rules and conventions used by the computer programs in annotating the tables. These rules and conventions were chosen based on feasibility considerations and a desire to balance statistical power with Type I error control within these feasibility constraints.

Two types of annotations were made. The first type of annotation ("<" or ">") was used to indicate a gain or loss that was statistically significant considering each jurisdiction (i.e., state, or territory) as a separate entity and controlling for the number of tests conducted in a particular table within that jurisdiction. Since all tables were set up with jurisdictions as the row variable, the first type of annotation was used on significance tests that *separately controlled the Type I error rate within each row of the table.* The second type of annotation ("<<" or ">>") was used to indicate a gain or loss that was statistically significant after *simultaneously controlling the Type I error rate for the number of tests conducted across all jurisdictions within a table.* As a result of this simultaneous error-rate control, the latter tests were extremely conservative and annotations of the second type were infrequent.

---

[8] Eighth-grade public-school results from the national assessment for the nation as a whole and for each region of the country are also shown in these tables. However, significance testing and table annotation was not carried out for these results. Statistical tests and annotations of differences for the national assessment were included in tables from the *Mathematics Report Card* that contain only national results.

Many of the tables contain two or more types of statistics. For example, a very common table in the *Data Compendium* contains, for both 1992 and 1990, the proportion of examinees in each of a particular set of reporting subgroups (e.g., males and females, or each of the race/ethnicity groups) *and* the average composite proficiency for each subgroup. In a table of this nature, two distinct families of significance tests were distinguished. The first family consisted of the comparisons of 1992 and 1990 proportions within each of the subgroups; the second consisted of the comparisons of 1992 and 1990 subgroup means. For each of these families, Type I error rates were controlled separately within-row (for the determining the first type of annotation) and simultaneously (for the second type of annotation).

As a second example, a different table contains the percentage of students in the top one-third of the schools, the average composite proficiency of these students, and the percentage of these students at or above each of the achievement levels. In this example, three families of significance tests were distinguished—tests comparing percentages in the top-third schools, tests comparing the average proficiencies of these students, and tests comparing the percentages exceeding the achievement levels. Again, Type I error rates were controlled separately within-row (for determining the first type of annotation) and simultaneously (for the second type of annotation) for each of these three families.

To illustrate the rules and conventions that were used, two specific examples will be considered. Table 10-4 is taken from an early version of Chapter 1 of the 1992 *Data Compendium.* It shows the 1992 average eighth-grade composite proficiencies at the percentages associated with each of the achievement levels for the nation, each of the four regions, and each state and territory that participated in the 1992 assessment. The same statistics are given for 1990, with "xxx" used for the 1992 participants that did not take part in the 1990 assessment. Two families of significance tests were distinguished for this table. The first family involved comparisons of 1992 average proficiencies with those obtained in 1990. The second family involved comparisons of 1992 and 1990 percentages at or above each of the achievement levels.

For the first family of tests (i.e., comparisons of average proficiencies), the annotations based on within-row control of Type I error required no Bonferroni adjustment. A t-test was carried out comparing each jurisdiction's 1992 average to the corresponding 1990 average at an α level of .05. For the annotation based on simultaneous control of Type I error, the family size was taken to be 37 (the 37 states and territories that participated in both the 1992 and 1990 assessments). In other words, the t-test for each jurisdiction was carried out at the $\alpha = .05/37$ level of significance.

For the second family of tests (i.e., comparisons of the percentage of students at or above each of the achievement levels), within-row control of Type I error *did* require adjustments to significance levels. Within each jurisdiction, three tests were carried out—one

TABLE 1.4 | Overall Average Mathematics Proficiency and Achievement Levels (continued)

PUBLIC SCHOOLS	Grade 8 - 1992					Grade 8 - 1990				
	Average Proficiency	Percentage of Students At or Above Advanced	Percentage of Students At or Above Proficient	Percentage of Students At or Above Basic	Percentage of Students Below Basic	Average Proficiency	Percentage of Students At or Above Advanced	Percentage of Students At or Above Proficient	Percentage of Students At or Above Basic	Percentage of Students Below Basic
NATION	266 (1.0)	3 (0.5)	23 (1.1)	61 (1.2)	39 (1.2)	262 (1.4)	2 (0.4)	19 (1.2)	57 (1.4)	43 (1.4)
Northeast	26[illegible] (3.0)	5 (1.4)	25 (3.0)	59 (3.9)	41 (3.9)	270 (3.3)	3 (1.0)	25 (3.1)	65 (3.7)	35 (3.7)
Southeast	258 (1.2)	1 (0.4)	16 (1.0)	53 (1.6)	47 (1.6)	254 (2.6)	2 (0.6)	15 (2.2)	48 (3.0)	52 (3.0)
Central	273 (2.2)	3 (0.7)	28 (3.0)	70 (2.8)	30 (2.8)	265 (2.3)	2 (0.6)	20 (2.1)	61 (2.5)	39 (2.5)
West	267 (2.1)	4 (1.1)	24 (2.1)	62 (2.7)	38 (2.7)	261 (2.6)	3 (0.7)	19 (2.5)	57 (2.6)	43 (2.6)
STATES										
Alabama	251 (1.7)	1 (0.3)	12 (1.1)	44 (2.0)	56 (2.0)	253 (1.1)	1 (0.2)	12 (0.8)	47 (1.6)	53 (1.6)
Arizona	265 (1.3) >	2 (0.4)	19 (1.4)	61 (1.8) >	39 (1.8) <	260 (1.3)	1 (0.4)	16 (1.1)	55 (1.8)	45 (1.8)
Arkansas	255 (1.2)	1 (0.3)	13 (1.0)	50 (1.7)	50 (1.7)	256 (0.9)	1 (0.2)	12 (1.0)	51 (1.3)	49 (1.3)
California	260 (1.7)	3 (0.7)	20 (1.4)	55 (2.0)	45 (2.0)	256 (1.3)	2 (0.4)	16 (1.3)	51 (1.6)	49 (1.6)
Colorado	272 (1.1) >	2 (0.5)	26 (1.3) >	69 (1.3) >	31 (1.3) <	267 (0.9)	2 (0.4)	22 (1.0)	64 (1.1)	36 (1.1)
Connecticut	273 (1.1) >	4 (0.6)	30 (1.1) >	69 (1.4)	31 (1.4)	270 (1.0)	4 (0.4)	26 (1.1)	66 (1.3)	34 (1.3)
Delaware	262 (1.0)	3 (0.4)	18 (1.1)	57 (1.2)	43 (1.2)	261 (0.9)	2 (0.5)	19 (0.9)	55 (1.3)	45 (1.3)
Dist. Columbia	234 (0.9) >	1 (0.2)	6 (1.0)	26 (1.3) >	74 (1.3) <	231 (0.9)	1 (0.2)	4 (0.7)	21 (1.0)	79 (1.0)
Florida	259 (1.5)	2 (0.4)	18 (1.3)	55 (1.9)	45 (1.9)	255 (1.3)	2 (0.4)	15 (1.0)	49 (1.4)	51 (1.4)
Georgia	259 (1.2)	1 (0.3)	16 (1.0)	53 (1.5)	47 (1.5)	259 (1.3)	3 (0.5)	17 (1.3)	53 (1.5)	47 (1.5)
Hawaii	257 (0.9)»	2 (0.4)	16 (0.8)	51 (1.2)»	49 (1.2)«	251 (0.8)	2 (0.4)	14 (0.8)	45 (1.0)	55 (1.0)
Idaho	274 (0.8) >	3 (0.4)	27 (1.2)	73 (1.1)	27 (1.1)	271 (0.8)	2 (0.4)	23 (1.4)	70 (1.2)	30 (1.2)
Indiana	269 (1.2)	3 (0.4)	24 (1.3)	66 (1.5)	34 (1.5)	267 (1.1)	3 (0.6)	21 (1.2)	63 (1.6)	37 (1.6)
Iowa	283 (1.0)»	5 (0.7)	37 (1.4) >	81 (1.2) >	19 (1.2) <	278 (1.1)	4 (0.5)	30 (1.5)	76 (1.1)	24 (1.1)
Kentucky	261 (1.1) >	2 (0.4)	17 (1.1)	57 (1.3) >	43 (1.3) <	257 (1.2)	1 (0.2)	14 (0.9)	51 (1.8)	49 (1.8)
Louisiana	249 (1.7)	1 (0.2)	10 (1.2)	42 (2.0)	58 (2.0)	246 (1.2)	1 (0.2)	8 (1.0)	39 (1.7)	61 (1.7)
Maine	278 (1.0)	4 (0.6)	31 (1.9)	77 (1.3)	23 (1.3)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)
Maryland	264 (1.3)	4 (0.6)	24 (1.3)	59 (1.5)	41 (1.5)	261 (1.4)	3 (0.6)	20 (1.2)	56 (1.7)	44 (1.7)
Massachusetts	272 (1.1)	3 (0.5)	28 (1.4)	68 (1.5)	32 (1.5)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)
Michigan	267 (1.4)	3 (0.5)	23 (1.7)	63 (1.6)	37 (1.6)	264 (1.2)	2 (0.4)	20 (1.4)	60 (1.4)	40 (1.4)
Minnesota	282 (1.0)»	6 (0.7) >	37 (1.2)»	79 (1.2) >	21 (1.2) <	275 (0.9)	4 (0.4)	29 (1.2)	74 (1.3)	26 (1.3)
Mississippi	246 (1.2)	0 (0.2)	8 (0.8)	38 (1.5)	62 (1.5)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)
Missouri	270 (1.2)	3 (0.4)	24 (1.3)	68 (1.6)	32 (1.6)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)
Nebraska	277 (1.1)	4 (0.5)	32 (1.9)	75 (1.2)	25 (1.2)	276 (1.0)	4 (0.6)	30 (1.4)	74 (1.1)	26 (1.1)
New Hampshire	278 (1.0)»	3 (0.6)	30 (1.5) >	77 (1.0) >	23 (1.0) <	273 (0.9)	3 (0.5)	25 (1.2)	71 (1.6)	29 (1.6)
New Jersey	271 (1.6)	4 (0.6)	28 (1.4)	67 (1.8)	33 (1.8)	270 (1.1)	4 (0.5)	25 (1.3)	65 (1.6)	35 (1.6)
New Mexico	259 (0.9) >	1 (0.3)	14 (1.0)	54 (1.4)	46 (1.4)	256 (0.7)	1 (0.3)	13 (0.9)	51 (1.3)	49 (1.3)
New York	266 (2.1)	4 (0.6)	24 (1.6) >	62 (2.3)	38 (2.3)	261 (1.4)	3 (0.5)	19 (1.0)	57 (1.7)	44 (1.7)
North Carolina	258 (1.2)»	1 (0.3)	15 (1.0) >	53 (1.5)»	47 (1.5)«	250 (1.1)	1 (0.4)	11 (0.8)	44 (1.4)	56 (1.4)
North Dakota	283 (1.2)	4 (0.6)	36 (1.7)	82 (1.3)	18 (1.3)	281 (1.2)	4 (0.6)	34 (2.0)	81 (1.6)	19 (1.6)
Ohio	267 (1.5)	2 (0.5)	22 (1.4)	64 (2.0)	36 (2.0)	264 (1.0)	2 (0.3)	19 (1.2)	60 (1.4)	40 (1.4)
Oklahoma	267 (1.2) >	2 (0.3)	21 (1.2) >	65 (2.0)	35 (2.0)	263 (1.3)	2 (0.5)	17 (1.3)	59 (1.6)	41 (1.6)
Pennsylvania	271 (1.5)	3 (0.7)	26 (1.5)	67 (1.7)	33 (1.7)	266 (1.6)	2 (0.4)	21 (1.5)	63 (2.0)	37 (2.0)
Rhode Island	265 (0.7)»	2 (0.3)	20 (1.3)	62 (1.2)»	38 (1.2)«	260 (0.6)	2 (0.3)	18 (1.0)	55 (0.9)	45 (0.9)
South Carolina	260 (1.0)	2 (0.5)	18 (1.1)	53 (1.2)	47 (1.2)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)
Tennessee	258 (1.4)	1 (0.4)	15 (1.2)	53 (1.8)	47 (1.8)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)
Texas	264 (1.3) >	4 (0.6)	21 (1.4) >	58 (1.5) >	42 (1.5) <	258 (1.4)	2 (0.4)	18 (1.0)	52 (1.7)	48 (1.7)
Utah	274 (0.7)	3 (0.5)	27 (1.1)	72 (1.3)	28 (1.3)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)
Virginia	267 (1.2)	3 (0.5)	23 (1.2)	62 (1.6)	38 (1.6)	264 (1.5)	4 (0.8)	21 (1.6)	58 (1.6)	42 (1.6)
West Virginia	258 (1.0)	1 (0.2)	13 (0.9)	53 (1.5)	47 (1.5)	256 (1.0)	1 (0.2)	12 (0.9)	49 (1.2)	51 (1.2)
Wisconsin	277 (1.5)	4 (0.6)	32 (1.4)	76 (1.9)	24 (1.9)	274 (1.3)	4 (0.5)	29 (1.5)	72 (1.7)	28 (1.7)
Wyoming	274 (0.9) >	2 (0.5)	26 (1.0)	73 (1.3)	27 (1.3)	272 (0.7)	2 (0.3)	24 (1.0)	71 (1.3)	29 (1.3)
TERRITORIES										
Guam	234 (1.0) >	1 (0.2)	7 (0.7)	30 (1.4)	70 (1.4)	232 (0.7)	1 (0.2)	5 (0.8)	27 (1.0)	73 (1.0)
Virgin Islands	222 (1.1) >	0 (0.1)	1 (0.3)	13 (1.0)	87 (1.0)	219 (0.9)	0 (0.1)	1 (0.4)	10 (1.1)	90 (1.1)

»The value for 1992 was significantly higher than the value for 1990 at about the 95 percent certainty level. «The value for 1992 was significantly lower than the value for 1990 at about the 95 percent certainty level. These notations indicate statistical significance from a muluple comparison procedure based on the 37 jurisdictions participating in both 1992 and 1990. If looking at only one state, then > and < also indicate differences that are significant. (xxx) Did not participate in the 1990 Trial State Assessment.

BEST COPY AVAILABLE

test for each of three of the achievement levels: At or Above Advanced, At or Above Proficient, and At or Above Basic[a]. Hence, the first type of annotations was based on t-tests carried out at the $\alpha = .05/3$ level. Across all jurisdictions there were 111 total comparisons (37 jurisdictions times 3 achievement levels). Hence, the annotations based on simultaneous control of Type I error were based on t-tests conducted at the $\alpha = .05/111$ level.

Table 10-5 is taken from an early version of Chapter 2 of the *Data Compendium*. For each jurisdiction it contains the 1992 and 1990 percentages of eighth-grade examinees in each race/ethnicity subgroup and, where subgroup sample sizes are 62 or greater, the 1992 and 1990 average composite proficiencies. Again, two families of significance tests were distinguished—tests comparing subgroup percentages and tests comparing subgroup means. For the first family of tests (i.e, comparisons of percentages within each race/ethnicity subgroup), five tests were carried out (one for each race/ethnicity group). Hence, the first type of annotation was based on t-tests carried out at the $\alpha = .05/5$ level and the second type of annotation was based on t-tests carried out at the $\alpha = .05/185$ (i.e., 37 jurisdictions times 5 race/ethnicity groups). For the second family of tests (i.e., comparisons of the subgroup average proficiencies), within-row control of Type I error required adjustments to significance levels based on the number of race/ethnicity groups with minimum sample sizes of 62. The annotations based on simultaneous control of Type I error required adjustments based on the number of subgroups across all 37 jurisdictions with minimum sample sizes of 62.

---

[a] Testing the percentage Below Basic is isomorphic to testing the percentage At or Above Basic. Therefore, it need not be counted as a distinct significance test.

TABLE 2.2 | Average Mathematics Proficiency by Race/Ethnicity (continued)

PUBLIC SCHOOLS	Grade 8 - 1992									
	White		Black		Hispanic		Asian / Pacific Islander		American Indian	
	Percentage of Students	Average Proficiency	Percentage of Students	Average Proficiency	Percentage of Students	Average Proficiency	Percentage of Students	Average Proficiency	Percentage of Students	Average Proficiency
NATION	69 (0.4)	276 (1.1)	16 (0.2)	236 (1.3)	10 (0.3)	245 (1.3)	2 (0.2)	287 (6.6)	1 (0.2)	254 (2.9)
Northeast	67 (2.6)	279 (3.3)	19 (1.5)	239 (3.8)	10 (1.7)	241 (3.8)!	2 (0.5)	*** (***)	1 (0.3)	*** (***)
Southeast	68 (1.8)	269 (1.2)	27 (1.8)	233 (1.7)	4 (0.7)	240 (2.8)!	1 (0.3)	*** (***)	1 (0.2)	*** (***)
Central	79 (2.0)	280 (2.0)	13 (1.6)	239 (3.5)	5 (0.8)	248 (4.2)	2 (0.5)	*** (***)	1 (0.4)	*** (***)
West	63 (1.5)	277 (2.4)	8 (1.3)	234 (3.5)	21 (1.7)	248 (1.6)	5 (0.8)	288(11.3)	2 (0.7)	*** (***)
STATES										
Alabama	61 (2.3)	264 (1.4)	32 (2.1)	231 (2.2)	4 (0.6)	220 (5.3)	1 (0.2)	*** (***)	2 (0.4)	*** (***)
Arizona	60 (2.1)	275 (1.1)	4 (0.5)	251 (3.4)	28 (1.6)	247 (2.7)	2 (0.3)	*** (***)	6 (1.3)	251 (2.7)»
Arkansas	72 (1.4)	265 (1.0)	22 (1.3)	230 (1.9)	4 (0.4)	228 (4.1)	1 (0.2)	*** (***)	1 (0.2)	*** (***)
California	44 (1.8)	276 (1.8)	7 (1.1)	233 (3.6)	36 (1.7)	240 (2.0)	11 (1.0)	276 (2.9)	1 (0.2)	*** (***)
Colorado	74 (1.2)	278 (1.0)	4 (0.6)	241 (4.4)	18 (1.1)	254 (1.7)>	2 (0.3)	*** (***)	2 (0.3)	*** (***)
Connecticut	72 (1.6)	283 (0.9)»	12 (1.1)	242 (2.9)	12 (0.9)	241 (2.4)	3 (0.4)	287 (6.0)	0 (0.1)	*** (***)
Delaware	65 (0.9)	272 (1.0)>	25 (1.1)	241 (1.8)	6 (0.6)	239 (3.4)	2 (0.3)	*** (***)	2 (0.3)	*** (***)
Dist. Columbia	3 (0.2)	*** (***)	85 (0.8)	233 (0.9)	10 (0.7)	225 (3.8)	1 (0.2)	*** (***)	1 (0.3)	*** (***)
Florida	56 (2.1)	273 (1.3)»	23 (2.0)	238 (2.3)	18 (2.0)	245 (2.5)	2 (0.3)	*** (***)	1 (0.2)	*** (***)
Georgia	59 (2.1)	270 (1.3)	35 (1.9)	241 (1.3)	4 (0.5)	233 (5.5)	2 (0.3)	*** (***)	0 (0.1)	*** (***)
Hawaii	17 (0.9)	265 (1.6)	3 (0.3)	*** (***)	11 (0.7)	238 (2.2)	66 (1.1)	258 (1.1)»	1 (0.2)	*** (***)
Idaho	88 (0.7)	277 (0.8)>	1 (0.2)	*** (***)	8 (0.6)	253 (2.3)	1 (0.2)	*** (***)	3 (0.4)	259 (4.2)
Indiana	85 (1.3)	273 (1.2)	8 (1.1)	243 (2.6)	4 (0.6)	249 (4.6)	1 (0.2)	*** (***)	1 (0.2)	*** (***)
Iowa	92 (0.7)	284 (1.0)>	2 (0.4)	*** (***)	4 (0.4)	261 (3.8)	1 (0.2)	*** (***)	1 (0.2)	*** (***)
Kentucky	87 (1.0)	264 (1.1)	9 (1.0)	241 (2.6)	3 (0.4)	231 (4.6)	1 (0.2)	*** (***)	1 (0.2)	*** (***)
Louisiana	54 (1.7)	253 (1.7)	39 (1.5)	232 (2.2)	5 (0.5)	228 (3.5)	2 (0.4)	*** (***)	1 (0.2)	*** (***)
Maine	94 (0.5)	279 (1.0)	0 (0.1)	*** (***)	2 (0.3)	*** (***)	1 (0.2)	*** (***)	3 (0.4)	261 (4.5)
Maryland	60 (1.8)	278 (1.5)	29 (1.8)	239 (2.0)	5 (0.6)	240 (3.3)	3 (0.5)	287 (4.7)	1 (0.2)	*** (***)
Massachusetts	83 (1.1)	277 (1.1)	5 (1.0)	243 (5.0)	8 (1.5)	240 (3.4)	2 (0.4)	*** (***)	1 (0.2)	*** (***)
Michigan	73 (1.8)	276 (1.5)>	18 (1.9)	232 (1.8)	5 (0.8)	248 (4.0)	1 (0.3)	*** (***)	2 (0.3)	*** (***)
Minnesota	91 (1.0)	284 (1.0)»	2 (0.3)	*** (***)	3 (0.5)	253 (3.8)	2 (0.3)	*** (***)	1 (0.4)	*** (***)
Mississippi	49 (1.9)	262 (1.4)	44 (1.8)	230 (1.4)	6 (0.6)	223 (3.1)	0 (0.1)	*** (***)	1 (0.2)	*** (***)
Missouri	82 (1.5)	275 (1.0)	12 (1.4)	241 (2.9)	3 (0.3)	251 (4.2)	1 (0.2)	*** (***)	2 (0.3)	*** (***)
Nebraska	87 (1.1)	281 (1.1)	5 (0.9)	236 (4.7)	6 (0.7)	254 (3.1)	1 (0.2)	*** (***)	2 (0.4)	*** (***)
New Hampshire	91 (1.6)	278 (0.9)>	1 (0.2)	*** (***)	3 (0.3)	258 (5.1)	1 (0.2)	*** (***)	1 (0.2)	*** (***)
New Jersey	61 (2.5)	283 (1.4)	17 (2.4)	242 (2.7)	14 (1.5)	247 (3.5)	6 (0.7)	297 (3.3)	1 (0.2)	*** (***)
New Mexico	44 (1.5)	272 (1.2)	2 (0.4)	*** (***)	49 (1.4)	248 (1.1)	1 (0.3)	*** (***)	4 (0.7)«	249 (3.0)>
New York	61 (2.7)	279 (1.1)>	17 (2.2)	232 (4.5)	14 (2.0)	243 (4.8)	4 (0.6)	281 (6.8)	1 (0.3)	*** (***)
North Carolina	68 (1.4)>	266 (1.0)>	27 (1.3)	238 (1.7)	3 (0.3)<	238 (4.7)>	1 (0.2)	*** (***)	2 (0.4)	*** (***)
North Dakota	93 (0.8)	284 (1.2)	0 (0.1)	*** (***)	3 (0.3)	*** (***)	1 (0.2)	*** (***)	3 (0.7)	261 (4.3)!»
Ohio	80 (1.9)	274 (1.4)>	14 (1.7)	234 (2.3)	4 (0.5)	245 (4.6)	1 (0.2)	*** (***)	2 (0.3)	*** (***)
Oklahoma	75 (1.6)	272 (1.0)	8 (1.1)	238 (3.0)	6 (0.6)	252 (3.2)	2 (0.3)	*** (***)	10 (1.0)	261 (3.2)
Pennsylvania	83 (1.4)	278 (1.1)	11 (1.6)	237 (4.6)	3 (0.7)	248 (3.9)!>	1 (0.3)	*** (***)	1 (0.3)	*** (***)
Rhode Island	81 (0.7)	271 (0.9)»	8 (0.6)	240 (2.9)>	8 (0.4)	232 (2.7)	3 (0.4)	254 (3.4)	2 (0.3)	*** (***)
South Carolina	58 (1.5)	273 (1.1)	35 (1.3)	241 (1.0)	6 (0.6)	233 (2.6)	1 (0.2)	*** (***)	1 (0.2)	*** (***)
Tennessee	75 (2.0)	266 (1.1)	21 (2.1)	234 (2.4)	3 (0.3)	227 (4.8)	0 (0.1)	*** (***)	1 (0.2)	*** (***)
Texas	48 (1.9)	279 (1.6)>	12 (1.6)	243 (2.0)>	36 (2.0)	248 (1.2)	3 (0.4)	301 (4.9)	1 (0.3)	*** (***)
Utah	90 (0.9)	276 (0.8)	1 (0.2)	*** (***)	7 (0.6)	253 (2.3)	2 (0.3)	*** (***)	2 (0.2)	*** (***)
Virginia	69 (1.9)	275 (1.1)	22 (1.6)	244 (1.9)	5 (0.6)	254 (4.0)	4 (0.5)	280 (4.0)	1 (0.2)	*** (***)
West Virginia	91 (0.9)	260 (1.0)	4 (0.8)	243 (3.7)	3 (0.3)<	230 (4.9)	0 (0.1)	*** (***)	2 (0.3)	*** (***)
Wisconsin	88 (1.7)	282 (1.2)	7 (1.2)	248 (6.8)	4 (0.6)	248 (4.0)	1 (0.2)<	*** (***)	2 (0.8)	261 (6.0)!
Wyoming	86 (1.7)	277 (0.8)	1 (0.2)	*** (***)	9 (0.8)	257 (2.1)	1 (0.2)	*** (***)	4 (1.6)	250 (2.4)!
TERRITORIES										
Guam	5 (0.5)<	266 (5.4)	1 (0.3)	*** (***)	15 (0.9)<	218 (2.8)	76 (1.1)>	236 (1.1)	1 (0.1)	*** (***)
Virgin Islands	1 (0.4)	*** (***)	77 (1.1)	224 (1.2)	21 (0.9)	213 (1.9)	0 (0.1)	*** (***)	0 (0.2)	*** (***)

»The value for 1992 was significantly higher than the value for 1990 at about the 95 percent certainty level. «The value for 1992 was significantly lower than the value for 1990 at about the 95 percent certainty level. These notations indicate statistical significance from a multiple comparison procedure based on the 37 jurisdictions participating in both 1992 and 1990. If looking at only one state, then > and < also indicate differences that are significant.

BEST COPY AVAILABLE

Table 10-5 (continued)

TABLE 2.2 | Average Mathematics Proficiency by Race/Ethnicity (continued)

PUBLIC SCHOOLS	Grade 8 - 1990									
	White		Black		Hispanic		Asian / Pacific Islander		American Indian	
	Percentage of Students	Average Proficiency	Percentage of Students	Average Proficiency	Percentage of Students	Average Proficiency	Percentage of Students	Average Proficiency	Percentage of Students	Average Proficiency
NATION	70 (0.5)	270 (1.5)	16 (0.3)	237 (2.8)	10 (0.4)	242 (2.8)	2 (0.5)	279 (5.4)!	2 (0.7)	244 (9.0)!
Northeast	80 (4.2)	274 (2.8)	12 (4.2)	246 (8.1)!	5 (1.2)	*** (****)	3 (1.1)	*** (****)	1 (0.3)	*** (****)
Southeast	63 (3.0)	265 (2.9)	32 (3.0)	235 (4.5)	3 (0.8)	*** (****)	1 (0.4)	*** (****)	0 (0.1)	*** (****)
Central	79 (2.8)	271 (2.4)	13 (3.2)	231 (5.2)!	5 (1.0)	*** (****)	1 (0.4)	*** (****)	1 (0.4)	*** (****)
West	63 (1.9)	269 (3.3)	7 (2.0)	245 (5.8)!	21 (1.5)	244 (3.4)	4 (1.3)	*** (****)	4 (2.3)	*** (****)
STATES										
Alabama	64 (1.9)	263 (1.0)	29 (1.8)	234 (1.8)	5 (0.6)	227 (3.7)	1 (0.3)	*** (****)	1 (0.2)	*** (****)
Arizona	59 (1.8)	271 (1.1)	3 (0.4)	245 (3.2)	29 (1.3)	242 (1.8)	2 (0.3)	*** (****)	7 (1.5)	235 (2.5)!
Arkansas	72 (1.5)	265 (0.9)	22 (1.5)	232 (1.2)	4 (0.4)	230 (4.0)	1 (0.2)	*** (****)	2 (0.3)	*** (****)
California	45 (1.8)	271 (1.5)	7 (0.8)	233 (3.4)	35 (1.4)	236 (1.8)	12 (1.1)	271 (2.8)	2 (0.4)	*** (****)
Colorado	73 (1.3)	274 (1.0)	4 (1.0)	237 (3.1)!	19 (1.8)	247 (1.4)	2 (0.3)	*** (****)	2 (0.3)	*** (****)
Connecticut	77 (1.5)	278 (0.9)	10 (1.0)	241 (2.4)	10 (0.9)	237 (2.7)	2 (0.3)	*** (****)	1 (0.2)	*** (****)
Delaware	68 (1.0)	268 (1.0)	24 (0.9)	242 (1.8)	5 (0.5)	242 (4.9)	1 (0.2)	*** (****)	1 (0.3)	*** (****)
Dist. Columbia	3 (0.4)	*** (****)	84 (1.0)	231 (0.7)	10 (0.8)	217 (3.1)	1 (0.2)	*** (****)	2 (0.3)	*** (****)
Florida	60 (2.0)	265 (1.4)	20 (1.2)	231 (1.7)	17 (2.1)	245 (2.8)	2 (0.4)	272 (5.1)	1 (0.2)	*** (****)
Georgia	59 (1.8)	271 (1.5)	33 (1.7)	240 (1.5)	6 (0.8)	231 (3.3)	1 (0.2)	*** (****)	1 (0.1)	*** (****)
Hawaii	18 (0.8)	263 (2.0)	2 (0.3)	*** (****)	10 (0.6)	231 (2.5)	67 (1.0)	252 (1.0)	1 (0.2)	*** (****)
Idaho	90 (0.8)	274 (0.8)	0 (0.1)	*** (****)	6 (0.6)	249 (2.8)	1 (0.3)	*** (****)	2 (0.4)	252 (4.8)
Indiana	84 (1.2)	271 (1.0)	9 (1.2)	243 (2.8)	4 (0.7)	245 (3.6)	1 (0.3)	*** (****)	1 (0.3)	*** (****)
Iowa	91 (0.7)	280 (1.1)	2 (0.7)	*** (****)	4 (0.4)	256 (3.9)	1 (0.2)	*** (****)	1 (0.3)	*** (****)
Kentucky	85 (1.1)	260 (1.2)	9 (1.0)	240 (2.4)	4 (0.5)	229 (3.5)	1 (0.2)	*** (****)	1 (0.2)	*** (****)
Louisiana	55 (2.1)	259 (1.4)	38 (1.9)	230 (1.3)	5 (0.6)	226 (4.2)	1 (0.2)	*** (****)	1 (0.3)	*** (****)
Maine	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)
Maryland	59 (1.5)	273 (1.5)	28 (1.5)	238 (1.8)	7 (0.8)	237 (2.9)	4 (0.7)	291 (4.3)	1 (0.3)	*** (****)
Massachusetts	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)
Michigan	77 (1.4)	271 (1.0)	13 (1.1)	232 (1.5)	5 (0.6)	243 (3.2)	2 (0.4)	*** (****)	2 (0.5)	*** (****)
Minnesota	90 (0.9)	278 (0.9)	2 (0.5)	239 (4.7)!	3 (0.4)	239 (5.0)	3 (0.4)	270 (5.8)	2 (0.5)	*** (****)
Mississippi	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)
Missouri	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)
Nebraska	88 (0.8)	279 (1.1)	5 (0.4)	235 (5.2)	5 (0.5)	253 (4.1)	1 (0.2)	*** (****)	1 (0.2)	*** (****)
New Hampshire	94 (0.5)	274 (0.9)	1 (0.2)	*** (****)	2 (0.4)	254 (4.2)	1 (0.2)	*** (****)	2 (0.2)	*** (****)
New Jersey	68 (2.0)	279 (1.2)	15 (2.0)	242 (2.3)	13 (1.0)	244 (2.2)	5 (0.8)	296 (4.3)	1 (0.2)	*** (****)
New Mexico	40 (1.3)	272 (1.2)	2 (0.4)	*** (****)	45 (1.3)	247 (1.1)	1 (0.3)	*** (****)	11 (0.8)	238 (1.4)
New York	60 (1.8)	274 (1.1)	17 (1.8)	236 (3.1)	17 (1.7)	237 (2.9)	4 (0.8)	278 (6.9)!	1 (0.3)	*** (****)
North Carolina	62 (1.7)	262 (1.3)	30 (1.3)	233 (1.3)	5 (0.5)	218 (3.3)	1 (0.2)	*** (****)	3 (0.9)	233 (4.3)!
North Dakota	91 (1.4)	284 (1.0)	1 (0.3)	*** (****)	3 (0.4)	248 (6.0)	1 (0.4)	*** (****)	5 (1.2)	242 (2.8)!
Ohio	82 (0.8)	269 (1.0)	11 (0.8)	233 (1.7)	3 (0.4)	237 (4.4)	1 (0.3)	*** (****)	1 (0.3)	*** (****)
Oklahoma	74 (1.8)	269 (1.3)	11 (1.2)	237 (2.2)	5 (0.7)	246 (4.3)	2 (0.4)	*** (****)	8 (1.0)	255 (2.5)
Pennsylvania	81 (2.5)	272 (1.1)	12 (2.3)	239 (3.1)	5 (0.8)	229 (4.5)	1 (0.2)	*** (****)	1 (0.3)	*** (****)
Rhode Island	83 (0.8)	266 (0.7)	5 (0.5)	227 (3.1)	8 (0.5)	230 (2.4)	2 (0.3)	*** (****)	1 (0.2)	*** (****)
South Carolina	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)
Tennessee	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)
Texas	47 (2.1)	273 (1.3)	13 (1.3)	236 (1.8)	36 (2.1)	245 (1.8)	2 (0.6)	*** (****)	1 (0.2)	*** (****)
Utah	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)	xxx (xxx)
Virginia	68 (1.5)	272 (1.8)	23 (1.5)	242 (1.8)	5 (0.5)	243 (4.1)	4 (0.4)	285 (4.2)	1 (0.2)	*** (****)
West Virginia	90 (0.7)	256 (0.9)	3 (0.5)	235 (4.1)	4 (0.4)	232 (4.2)	1 (0.2)	*** (****)	2 (0.3)	*** (****)
Wisconsin	85 (1.2)	279 (1.1)	8 (1.1)	237 (4.2)	4 (0.3)	250 (3.8)	2 (0.3)	*** (****)	1 (0.2)	*** (****)
Wyoming	86 (0.8)	275 (0.7)	1 (0.2)	*** (****)	8 (0.8)	255 (2.2)	1 (0.2)	*** (****)	3 (0.4)	257 (3.4)
TERRITORIES										
Guam	7 (0.7)	257 (3.5)	1 (0.4)	*** (****)	19 (1.0)	210 (1.9)	72 (1.2)	235 (0.9)	1 (0.2)	*** (****)
Virgin Islands	2 (0.2)	*** (****)	77 (1.1)	221 (1.1)	20 (1.0)	209 (1.5)	0 (0.2)	*** (****)	1 (0.2)	*** (****)

(xxx) Did not participate in the 1990 Trial State Assessment.

APPENDIX A

PARTICIPANTS IN THE OBJECTIVES AND ITEM DEVELOPMENT PROCESS

APPENDIX A

PARTICIPANTS IN THE OBJECTIVES AND ITEM DEVELOPMENT PROCESS

The National Assessment of Educational Progress extends its deep appreciation to all those individuals who participated in the development of the framework, objectives, and items for the Trial State Assessment Program in mathematics.

NATIONAL ASSESSMENT PLANNING PROJECT

Steering Committee

Robert Astrup	National Education Association
Lillian Barna	Council of the Great City Schools
Richard A. Boyd	Council of Chief State School Officers
Glenn Bracht	Council for American Private Education and National Association of Independent Schools
William M. Ciliate	National School Boards Association
Antonia Cortese	American Federation of Teachers
Mary Brian Costello	National Community on Catholic Education Association
Wilhelmina Delco	National Council of State Legislators
Nancy DiLaura	National Governors' Association
Thomas Fisher	Association of State Assessment Programs
Alice Houston	Association for Supervision and Curriculum Development
C. June Knight	National Association of Elementary School Principals
Stephen Lee	National Association of Secondary School Principals
Paul LeMahieu	National Association of Test Directors
Glen Ligon	Directors of Research and Evaluation
Barbara Roberts Mason	National Association of State Boards of Education
James E. Morrell	American Association of School Administrators, Austin Independent School District, Texas

Mathematics Objectives Committee

Joan Burks	Damascus High School, Damascus, Maryland
Phillip Curtis	University of California at Los Angeles, Los Angeles, California

Walter Denham	California Department of Education, Sacramento, California
Thomas Fisher	Florida Department of Education, Tallahassee, Florida
Ann Kahn	The National Parent-Teacher Association, Fairfax, Virginia
Mary M. Lindquist	Columbus College, Columbus, Georgia
Susan Purser	Whitten Junior High School, Jackson, Mississippi
Dorothy Strong	Chicago Public Schools, Chicago, Illinois
Thomas W. Tucker	Colgate University, Hamilton, New York
Charles Watson	Arkansas Department of Education, Little Rock, Arkansas
O.R. Wells, Jr.	Rice University, Houston, Texas

NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

Item Development Panel

Martha Baca	Roosevelt School District, Phoenix, Arizona
Iris Carl	Houston Independent School District, Houston, Texas
Carol Cho	Alhambra High School, Martinez, California
John Dossey	Illinois State University, Normal, Illinois
Elizabeth Fennema	University of Wisconsin, Madison, Wisconsin
Steven Leinwand	Connecticut State Department of Education, Middletown, Connecticut
Mary M. Lindquist	Columbus College, Columbus, Georgia
Gloria Sanok	Packanack School, West Caldwell, New Jersey
Thomas Tucker	Colgate University, Hamilton, New York

Test Development Consultants

James Braswell	College Board Programs, Educational Testing Service
Gail Chapman	College Board Programs, Educational Testing Service
Jeanne Galick	College Board Programs, Educational Testing Service
Jeffrey Haberstroh	College Board Programs, Educational Testing Service
Chancey Jones	College Board Programs, Educational Testing Service
Patricia Klag	College Board Programs, Educational Testing Service

Jane Kupin	College Board Programs, Educational Testing Service
Jane Maroney	College Board Programs, Educational Testing Service
Marlene Supernavage	College Board Programs, Educational Testing Service
Beverly Whittington	College Board Programs, Educational Testing Service

Scale Anchoring Panel

Laurie Boswell	Profile High School, Bethlehem, New Hampshire
Bruce Brombacher	Jones Junior High School, Westerville, Ohio
Catherine Brown	Virginia Polytechnic Institute, Blacksburg, Virginia
Joe Crosswhite	Springfield, Missouri
John Dossey	Illinois State University, Normal, Illinois
Henry Kepner, Jr.	University of Wisconsin at Milwaukee, Milwaukee, Wisconsin
Linda Kolnowski	Detroit Public Schools, Detroit, Michigan
Gordon Lewis	Washington DC Public Schools, Washington, D.C.
Mary M. Lindquist	Columbus College, Columbus, Georgia
Donna Long	Indiana Dept. of Education, Indianapolis, Indiana
Vena Long	Missouri Department of Education, Jefferson City, Missouri
William Masalski	University of Massachusetts, Amherst, Massachusetts
Wendell Meeks	Illinois State Board of Education, Springfield, Illinois
Andy Reeves	Florida Department of Education, Tallahassee, Florida
Diane Thiessen	University of Northern Iowa, Cedar Falls, Iowa
Alba Thompson	CRMSE, San Diego, CA 92120
Shiela Vice	Kentucky Department of Education, Frankfort, Kentucky
Charles Watson	Arkansas Department of Education, Little Rock, Arkansas
Vernon Williams	H.W. Longfellow Intermediate School, Falls Church, Virginia

APPENDIX B

SUMMARY OF PARTICIPATION RATES

Guidelines for Sample Participation and Explanation of Derivation of Weighted Participation

Introduction

Since 1989, state representatives, the National Assessment Governing Board (NAGB), several committees of external advisors to the National Assessment of Educational Progress (NAEP), and the National Center for Education Statistics (NCES) have engaged in numerous discussions about the procedures for reporting the NAEP Trial State Assessment results. As part of these discussions, it was recognized that sample participation rates across the states and territories have to be uniformly high to permit fair and valid comparisons. Therefore, NCES established four guidelines for school and student participation in the 1990 Trial State Assessment Program.

The participation rate data were presented in the appendix of the 1990 composite mathematics report (*The State of Mathematics Achievement*) and a notation was made in those appendix tables and in Table 2 of the appropriate state report for any jurisdiction with participation levels that did not meet the guidelines. Virtually every state and territory met or exceeded the four guidelines for the 1990 program.

For the 1992 Trial State Assessment, NCES has decided to continue to use those four guidelines, two relating to school participation and two relating to student participation. The guidelines are based on the standards for sample surveys that are set forth in the U.S. Department of Education's *Standards and Policies* (1987). Three of the guidelines for the 1992 program are identical to those used in 1990, while one guideline for school participation has been modified.

NCES and NAGB have reviewed the policy of how participation rates can best be presented so that readers of reports can accurately assess the quality of the data being reported. They have decided that for reporting the results from the 1992 Trial State Assessment Program, tables again will have notations for the jurisdictions not meeting each guideline. They also have decided that there will be a fuller discussion in the body of the 1992 composite reports about the participation rates and nature of the samples for each of the participating jurisdictions.

The next section of this report provides an explanation of the guidelines and notations. In brief, the guidelines cover levels of school and student participation, both overall and for particular population classes. Consistent with the NCES standards, weighted data must be used to calculate all participation rates for sample surveys, and weighted rates will be provided in the reports. The procedures used to derive the weighted school and student participation rates are provided immediately following the discussion of the guidelines and notations.

The final section of this report consists of a set of tables that provide the 1992 participation rate information for the 1992 Trial State mathematics assessment. Because the aggregate across all states is not representative of any meaningful sample, the weighted participation rates across states have not been analyzed. However, the national and regional counts from the national assessment have been included and do provide some context for interpreting the summary of activities in each individual state and territory.

Notations for Use in Reporting School and Student Participation Rates

Unless the overall participation rate is high for a state or territory, there is a risk that the assessment results for that jurisdiction are subject to appreciable nonresponse bias. Moreover, even if the overall participation rate is high, there may be significant nonresponse bias if the nonparticipation that does occur is heavily concentrated among certain classes of schools or students.

The following notations concerning school and student participation rates in the Trial State Assessment Program were established to address four significant ways in which nonresponse bias could be introduced into the jurisdiction sample estimates. The four conditions that result in a state or territory receiving a notation in the 1992 reports are presented below. Note that in order to receive no notations, a state or territory must satisfy all the guidelines at both grade 4 and grade 8.

A jurisdiction will receive a notation if:

1. **Both the state's weighted participation rate for the initial sample of schools was below 85 percent AND the weighted school participation rate after substitution was below 90 percent; OR the weighted school participation rate of the initial sample of schools was below 70 percent (regardless of the participation rate after substitution.)**

Discussion: For states or territories that did not use substitute schools, the participation rates are based on participating schools from the original sample. In these situations, the NCES standards specify weighted school participation rates of at least 85 percent to guard against potential bias due to school nonresponse. Thus, the first part of the notation that refers to the weighted school participation rate for the initial sample of schools is in direct accordance with NCES standards.

To help ensure adequate sample representation for each jurisdiction participating in the 1992 Trial State Assessment Program, NAEP provided substitutes for nonparticipating schools. When possible, a substitute school was provided for each initially selected school that declined participation before November 15, 1991. For states or territories that used substitute schools, the assessment results will be based on the student data from all participating schools from both the original sample and the list of substitutes (unless both an initial school and its substitute eventually participated, in which case only the data from the initial school will be used).

The NCES standards do not explicitly address the use of substitute schools to replace initially selected schools that decide not to participate in the assessment. However, considerable technical consideration was given to this issue. Even though the characteristics of the substitute schools were matched as closely as possible to the characteristics of the initially selected schools, substitution does not entirely eliminate bias due to the nonparticipation of initially selected schools. Thus, for the weighted school participation rates including substitute schools, the guideline was set at 90 percent.

Finally, if the jurisdiction's school participation rate for the initial sample of schools is below 70 percent, even if the rate after substitution exceeds 90 percent, there is a substantial possibility that, in aggregate, the substitute schools are not sufficiently similar to the schools that they replaced to assure that there is negligible bias in the assessment results. The last part of the notation takes this into consideration.

A jurisdiction will receive a notation if:

2. **The nonparticipating schools included a class of schools with similar characteristics, which together accounted for more than five percent of the state's total fourth- or eighth-grade weighted sample of public schools. The classes of schools from each of which a state needed minimum school participation levels were determined by urbanicity, minority enrollment, and median household income of the area in which the school is located.**

Discussion: The NCES standards specify that attention should be given to the representativeness of the sample coverage. Thus, if some important segment of the jurisdiction's population is not adequately represented, it is of concern, regardless of the overall participation rate.

This notation addresses the fact that, if nonparticipating schools are concentrated within a particular class of schools, the potential for substantial bias remains, even if the overall level of school participation appears to be satisfactory. Nonresponse adjustment cells have been formed within each jurisdiction, and the schools within each cell are similar with respect to minority enrollment, urbanicity, and/or median household income, as appropriate for each jurisdiction.

If more than five percent (weighted) of the sampled schools (after substitution) are nonparticipants from a single adjustment cell, then the potential for nonresponse bias is too great. This guideline is based on the NCES standard for stratum-specific school nonresponse rates.

A jurisdiction will receive a notation if:

3. **The weighted student response rate within participating schools was below 85 percent.**

Discussion: This guideline follows the NCES standard of 85 percent for overall student participation rates. The weighted student participation rate is based on all eligible students from initially selected or substitute schools who participated in the assessment in either an initial session or a make-up session. If the rate falls below 85 percent, then the potential for bias due to students' nonresponse is too great.

A jurisdiction will receive a notation if:

4. **The nonresponding students within participating schools included a class of students with similar characteristics, who together comprised more than five percent of the state's weighted assessable student sample. Student groups from which a state needed minimum levels of participation were determined by age of student and type of assessment session (unmonitored or monitored), as well as school urbanicity, minority enrollment, and median household income of the area in which the school is located.**

Discussion: This notation addresses the fact that if nonparticipating students are concentrated within a particular class of students, the potential for substantial bias remains, even if the overall student participation level appears to be satisfactory. Student nonresponse adjustment cells have been formed using the school-level nonresponse adjustment cells, together with the student's age and the nature of the assessment session (unmonitored or monitored). If more than five percent (weighted) of the invited students who do not participate in the assessment are from a single adjustment cell, then the potential for nonresponse bias is too great. This guideline is based on the NCES standard for stratum-specific student nonresponse rates.

Derivation of Weighted Participation Rates

Weighted School Participation Rates. The weighted school participation rates within each state or territory provide the percentages of fourth- or eighth-grade students in public schools who are represented by the schools participating in the assessment, prior to statistical adjustments for school nonresponse.

Two weighted school participation rates are computed per subject per grade for each state and territory. The first is the weighted participation rate for the initial sample of schools. This rate is based only on those schools that were initially selected for the assessment. The numerator of this rate is the sum of the number of students represented by each initially selected school that participated in the assessment. The denominator is the sum of the number of students represented by each of the initially selected schools found to have eligible students enrolled. This includes both participating and nonparticipating schools.

The second participation rate is the weighted participation rate after substitution. The numerator of this rate is the sum of the number of students represented by each of the participating schools, whether originally selected or a substitute. The denominator is the same as that for the weighted participation rate for the initial sample. This means that, for a given state, grade, and subject, the weighted participation rate after substitution is always at least as great as the weighted participation rate for the initial sample of schools.

In general, different schools in the sample can represent different numbers of students in the state population. The number of students represented by an initially selected school (the school weight) is the fourth- or eighth-grade enrollment of the school divided by the probability that the school was included in the sample. For instance, a selected school with an eighth-grade enrollment of 150 and a selection probability of 0.2 represents 750 students from that state. The number of students represented by a substitute school is the number of students represented by the replaced nonparticipating school.

Because each selected school represents different numbers of students in the population, the weighted school participation rates may differ somewhat from the simple unweighted rates. (The unweighted rates are calculated from the counts of schools by dividing the number of participating schools by the number of schools in the sample.) The difference between the weighted and the unweighted rates is potentially largest in smaller jurisdictions where all schools with fourth- or eighth-grade students were included in the sample. In those jurisdictions, each school represents only its own students. Therefore, the nonparticipation of a large school reduces the weighted school participation rate by a greater amount than does the nonparticipation of a small school.

The nonparticipation of larger schools also has greater impact than that of smaller schools on reducing weighted school participation rates in larger jurisdictions where fewer than all of the schools were included in the sample. However, since the number of students represented by each school is more nearly constant in larger states, the difference between the impact of nonparticipation by either large or small schools is less marked than in states where all schools were selected.

In general, the larger the jurisdiction, the less the difference is between the weighted and unweighted school participation rates. However, even in the smaller jurisdictions, the differences tend to be small.

Weighted Student Participation Rate. The weighted student participation rate provides the percentage of the eligible student population from participating schools within the state or territory that are represented by the students who participated in the assessment (in either an initial session or a make-up session). The eligible student population from participating schools within a jurisdiction consists of all public-school students who were in the fourth grade or eighth grade, who attended a school that, if selected, would have participated and who, if selected, would not have been excluded from the assessment. The numerator of this rate is the sum, across all assessed students, of the number of students represented by each assessed student (prior to adjustment for student nonparticipation). The denominator is the sum of the number of students represented by each selected student who was invited and eligible to participate (i.e., not excluded), including students who did not participate. Thus, the denominator is an estimate of the total number of assessable students in the group of schools within the jurisdiction that would have participated if selected.

The number of students represented by a single selected student (the student weight) is 1.0 divided by the overall probability that the student was selected for assessment. In general, the number of students from a jurisdiction's population that are represented by a sampled student is approximately constant across students. Consequently, there is little difference between the weighted student participation rate and the unweighted student participation rate.

Weighted Overall School and Student Participation Rate. An overall indicator of the effect of nonparticipation by both students and schools is given by the overall participation rate. This is calculated as the product of the weighted school participation rate (after substitution), and the weighted student participation rate. For jurisdictions having a high overall participation rate the potential is low for bias to be introduced through either school nonparticipation or student nonparticipation. This rate provides a summary measure that indicates the proportion of the jurisdiction's fourth- or eighth-grade student population that is directly represented by the final student sample. When the overall rate is high, the adjustments for nonresponse that are used in deriving the final survey weights are likely to be effective in maintaining nonresponse bias at a negligible level. Conversely, when the overall rate is relatively low there is a greater chance that a nonnegligible bias remains even after making such adjustments.

The overall rate is not used in establishing the guidelines/notations for school and student participation, since guidelines already exist covering school and student participation separately. The overall participation rate was not reported in 1990.

Derivation of Weighted Percentages for Excluded Students

Weighted Percentage of Excluded Students. The weighted percentage of excluded students estimates the percentage of the fourth- or eighth-grade population in the jurisdiction's public schools that are represented by the students who were excluded from the assessment, after accounting for school nonparticipation. The numerator is the sum, across all excluded students, of the number of students represented by each excluded student. The denominator is the sum of the number of students represented by each of the students who was sampled (and had not withdrawn from the school at the time of the assessment).

Weighted Percentage of Students with an Individualized Education Plan (IEP). The weighted percentage of IEP students estimates the percentage of the fourth- or eighth-grade population in the jurisdiction's public schools that are represented by the students who were classified as IEP, after accounting for school nonparticipation. The numerator is the sum, across all students classified as IEP, of the number of students represented by each IEP student. The denominator is the sum of the number of students represented by each of the students who was sampled (and had not withdrawn from the school at the time of the assessment).

Weighted Percentage of Excluded IEP Students. The weighted percentage of IEP students who were excluded estimates the percentage of students in the jurisdiction that are represented by those IEP students who were excluded from the assessment, after accounting for school nonparticipation. The numerator is the sum, across all students classified as IEP and excluded from the assessment, of the number of students represented by each excluded IEP student. The denominator is the sum of the number of students represented by each of the IEP students who was sampled (and had not withdrawn from the school at the time of the assessment).

Weighted Percentage of Limited English Proficiency (LEP) Students. The weighted percentage of LEP students estimates the percentage of the fourth- or eighth-grade population in the jurisdiction's public schools that are represented by the students who were classified as LEP, after accounting for school nonparticipation. The numerator is the sum, across all students classified as LEP, of the number of students represented by each LEP student. The denominator is the sum of the number of students represented by each of the students who was sampled (and had not withdrawn from the school at the time of the assessment).

Weighted Percentage of Excluded LEP Students. The weighted percentage of LEP students who were excluded estimates the percentage of students in the jurisdiction that are represented by those LEP students who were excluded from the assessment, after accounting for school nonparticipation. The numerator is the sum, across all students classified as LEP and excluded from the assessment, of the number of students represented by each excluded LEP student. The denominator is the sum of the number of students represented by each of the LEP students who was sampled (and had not withdrawn from the school at the time of the assessment).

TABLE B.1 | School Participation Rates

PUBLIC SCHOOLS	Grade 4 - 1992							
	Weighted Percentage School Participation Before Substitution	Weighted Percentage School Participation After Substitution	Number Schools in Original Sample	Number Schools Not -Eligible	Number Schools in Original Sample that Participated	Number Substitute Schools Provided	Number Substitute Schools that Participated	Total Number Schools that Participated
NATION	86	88	313	4	268	7	2	270
Northeast	82	82	59	0	49	1	0	49
Southeast	84	84	81	1	76	1	1	77
Central	82	82	88	1	82	0	0	82
West	79	79	105	2	81	5	1	82
STATES								
Alabama	75	87	113	3	81	27	25	106
Arizona	100	100	110	2	108	0	0	108
Arkansas[4]	90	99	123	2	109	11	11	120
California	91	97	115	3	101	7	7	108
Colorado	100	100	123	2	121	0	0	121
Connecticut	99	99	115	4	110	0	0	110
Delaware[2,3]	88	88	56	6	44	0	0	44
Dist. Columbia	99	99	114	6	107	0	0	107
Florida	100	100	111	1	110	0	0	110
Georgia	100	100	110	2	108	0	0	108
Hawaii	100	100	108	0	108	0	0	108
Idaho	84	97	120	0	98	21	17	115
Indiana	76	91	118	2	88	26	17	105
Iowa	100	100	132	4	128	0	0	128
Kentucky[4]	93	96	124	3	115	3	3	118
Louisiana	100	100	113	4	109	0	0	109
Maine[1,2,4,5]	57	71	142	2	75	44	23	98
Maryland	99	99	112	1	110	1	0	110
Massachusetts	87	97	123	4	103	12	11	114
Michigan[4]	83	90	114	3	90	18	8	98
Minnesota[4]	82	94	118	5	93	16	14	107
Mississippi	96	100	111	2	107	2	2	109
Missouri	88	97	120	7	101	9	9	110
Nebraska[1,2]	80	87	157	8	109	36	11	120
New Hampshire[1,2,4,5]	69	80	126	3	84	38	20	104
New Jersey[1,2]	76	82	119	3	88	22	7	95
New Mexico[4,5]	75	90	116	2	86	30	22	108
New York[1,2,4]	78	83	107	0	83	21	7	90
North Carolina[4]	95	99	118	2	111	5	5	116
North Dakota	73	90	133	3	97	30	19	116
Ohio	79	91	122	1	95	21	15	110
Oklahoma	86	96	129	3	111	14	13	124
Pennsylvania	84	95	116	0	99	17	12	111
Rhode Island	83	96	115	5	90	15	15	105
South Carolina	96	99	112	2	108	1	1	109
Tennessee	92	93	120	2	108	8	1	109
Texas	93	96	111	3	100	5	5	105
Utah	99	99	110	1	108	0	0	108
Virginia	99	99	116	4	111	0	0	111
West Virginia	100	100	147	6	141	0	0	141
Wisconsin	100	100	127	5	122	0	0	122
Wyoming	97	97	157	11	143	0	0	143
TERRITORIES								
Guam[2,3]	94	94	21	0	20	0	0	20
Virgin Islands[3]	100	100	24	0	24	0	0	24

See explanations of the notations and guidelines about sample representativeness and for the derivation of weighted participation. Weighted percentages for the nation and region are based on schools sampled for all subject areas assessed in 1990 (reading, science, and mathematics) or 1992 (mathematics, reading, and writing). However, based on the national sampling design, the rates shown also are the best estimates for the mathematics assessment. [1]Both the state's weighted participation rate for the initial sample of schools was below 85% AND the weighted school participation rate after substitution was below 90%; OR the weighted school participation rate of the initial sample of schools was below 70% (regardless of the participation rate after substitution.) [2]The nonparticipating schools included a class of schools with similar characteristics, which together accounted for more than five percent of the state's total fourth- or eighth-grade weighted sample of public schools. The classes of schools from each of which a state needed minimum school participation levels were determined by urbanicity, minority enrollment, and median household income of the area in which the school is located.

BEST COPY AVAILABLE

TABLE B.1 | School Participation Rates (continued)

PUBLIC SCHOOLS	Grade 8 - 1992							
	Weighted Percentage School Participation Before Substitution	Weighted Percentage School Participation After Substitution	Number Schools in Original Sample	Number Schools Not Eligible	Number Schools in Original Sample that Participated	Number Substitute Schools Provided	Number Substitute Schools that Participated	Total Number Schools that Participated
NATION	88	89	248	1	216	4	3	219
Northeast	92	92	45	0	41	0	0	41
Southeast	94	94	62	0	57	0	0	57
Central	86	87	61	1	52	1	1	53
West	82	84	80	0	66	3	2	68
STATES								
Alabama[1] [4]	66	92	107	1	70	37	32	102
Arizona[4]	99	99	108	5	103	0	0	103
Arkansas[4]	89	97	101	1	89	10	8	97
California[4]	93	98	107	2	98	7	6	104
Colorado	100	100	113	1	112	0	0	112
Connecticut	98	99	101	3	97	0	0	97
Delaware[3]	100	100	30	2	28	0	0	28
Dist. Columbia[3]	100	100	37	2	35	0	0	35
Florida	100	100	107	4	103	0	0	103
Georgia[4]	99	99	106	4	102	0	0	102
Hawaii[3]	100	100	57	5	51	0	0	51
Idaho[2] [4]	85	91	82	1	67	12	7	74
Indiana[4]	79	94	107	0	85	21	17	102
Iowa	99	99	109	3	105	0	0	105
Kentucky	98	98	112	6	102	3	2	104
Louisiana	100	100	109	8	101	0	0	101
Maine[1] [2] [4]	82	84	100	0	60	33	22	82
Maryland[4]	89	91	104	1	93	9	2	95
Massachusetts	83	95	109	7	85	12	12	97
Michigan	78	94	108	1	83	22	18	101
Minnesota	81	92	104	3	82	15	11	93
Mississippi	99	100	102	3	98	1	1	99
Missouri	92	99	107	1	98	7	7	105
Nebraska[1]	75	85	122	10	73	34	12	85
New Hampshire[4]	80	92	78	1	62	14	11	73
New Jersey[1] [2]	69	78	108	2	75	27	9	84
New Mexico[4]	77	94	93	1	69	22	16	85
New York[1] [2]	81	83	108	4	84	19	3	87
North Carolina	94	98	108	3	99	4	4	103
North Dakota	78	97	80	6	55	16	15	70
Ohio[4]	77	90	110	0	85	20	14	99
Oklahoma	82	96	110	3	88	17	17	105
Pennsylvania[4]	81	94	107	2	84	21	15	99
Rhode Island[3]	85	100	57	5	44	7	7	51
South Carolina	94	97	105	0	99	4	3	102
Tennessee	87	91	108	2	91	10	4	95
Texas	95	99	107	3	99	5	4	103
Utah	100	100	88	3	85	0	0	85
Virginia	97	97	108	2	103	0	0	103
West Virginia	100	100	108	4	104	0	0	104
Wisconsin	100	100	109	2	107	0	0	107
Wyoming	99	99	66	11	54	0	0	54
TERRITORIES								
Guam[3]	100	100	6	0	6	0	0	6
Virgin Islands[3]	100	100	6	0	6	0	0	6

[3] The Trial State Assessment was based on all eligible schools. There was no sampling of schools. [4] In one or more schools an assessment was conducted, but either the wrong materials were sent to the school(s) or the materials were lost in shipping via the U.S. Postal Service. The school(s) are included in the counts of participating schools, both before and after substitution. However, in the weighted results, the school(s) are treated in the same manner as a nonparticipating school because no student responses were available for analysis and reporting.

BEST COPY AVAILABLE

TABLE B.1 | **School Participation Rates (continued)**

	Grade 8 - 1990							
PUBLIC SCHOOLS	Weighted Percentage School Participation Before Substitution	Weighted Percentage School Participation After Substitution	Number Schools in Original Sample	Number Schools Not Eligible	Number Schools in Original Sample that Participated	Number Substitute Schools Provided	Number Substitute Schools that Participated	Total Number Schools that Participated
NATION	88	92	145	13	117	15	3	120
Northeast	72	90	25	3	17	5	2	19
Southeast	94	94	40	1	35	4	0	35
Central	94	94	31	4	26	1	0	26
West	88	90	49	5	39	5	1	40
STATES								
Alabama	86	97	106	5	87	13	11	98
Arizona[4]	97	97	110	7	102	0	0	102
Arkansas	100	100	107	0	107	0	0	107
California	94	94	106	2	98	0	0	98
Colorado	100	100	107	2	105	0	0	105
Connecticut	100	100	108	5	103	0	0	103
Delaware[3]	100	100	30	0	30	0	0	30
Dist. Columbia[3]	100	100	36	0	36	0	0	36
Florida[4]	98	98	108	6	101	0	0	101
Georgia	100	100	109	3	106	0	0	106
Hawaii[3]	100	100	57	4	52	0	0	52
Idaho	97	97	108	2	101	4	0	101
Indiana[4]	89	94	105	1	92	9	6	98
Iowa[3]	91	91	108	7	92	9	0	92
Kentucky	100	100	112	8	104	0	0	104
Louisiana	100	100	108	9	99	0	0	99
Maine	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
Maryland	100	100	107	2	105	0	0	105
Massachusetts	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
Michigan	90	97	105	4	90	9	8	98
Minnesota	90	93	108	3	94	5	3	97
Mississippi	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
Missouri	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
Nebraska	87	94	121	8	94	10	9	103
New Hampshire	91	97	107	3	94	4	4	98
New Jersey	97	98	112	3	106	2	1	107
New Mexico	100	100	108	2	106	0	0	106
New York[4]	86	86	105	0	91	0	0	91
North Carolina	100	100	111	5	106	0	0	106
North Dakota	96	100	111	5	98	8	8	106
Ohio	96	98	105	2	99	4	2	101
Oklahoma	78	99	112	0	85	26	23	108
Pennsylvania	90	93	106	4	92	4	3	95
Rhode Island[3]	94	97	52	0	49	2	2	51
South Carolina	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
Tennessee	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
Texas[3]	88	97	107	4	92	10	9	101
Utah	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
Virginia	99	99	106	1	104	0	0	104
West Virginia	100	100	107	6	101	0	0	101
Wisconsin[4]	99	99	109	3	106	0	0	106
Wyoming	100	100	69	0	69	0	0	69
TERRITORIES								
Guam[3]	100	100	7	1	6	0	0	6
Virgin Islands[3]	100	100	6	0	6	0	0	6

[3] One or more schools in the original sample initially declined and then decided to participate after their substitute(s) had also agreed to participate. Further, assessments were conducted in both the original and substitute schools. For these cases the substitute school is included in the number of substitute schools provided and in the number of substitute schools participating. The state's estimates will be based on the student responses from the original school only. (xxx) Did not participate in the 1990 Trial State Assessment.

TABLE B.2 | Student Participation Rates

PUBLIC SCHOOLS	Grade 4 - 1992								
	Weighted Percentage Student Participation After Make-ups	Number Students Original Sample	Number Students Supplemental Sample	Number Students Withdrawn	Number Students Excluded	Number Students to be Assessed	Number Students Assessed Initial Sessions	Number Students Assessed Make-ups	Total Number Students Assessed
NATION	94	6,582	–	–	584	5,998	5,839	2	5,841
Northeast	94	1,175	–	–	106	1,069	1,007	0	1,007
Southeast	93	1,981	–	–	133	1,848	1,733	0	1,733
Central	94	1,357	–	–	75	1,282	1,213	0	1,213
West	94	2,069	–	–	270	1,799	1,686	2	1,688
STATES									
Alabama	95	2,903	68	115	127	2,729	2,605	0	2,605
Arizona	95	3,133	152	232	154	2,899	2,752	10	2,762
Arkansas[4]	96	2,961	90	149	154	2,748	2,641	6	2,647
California	94	3,015	141	224	364	2,568	2,392	20	2,412
Colorado	95	3,244	124	152	166	3,050	2,893	13	2,906
Connecticut	96	2,959	68	118	196	2,713	2,596	4	2,600
Delaware	95	2,330	84	141	121	2,152	2,028	12	2,040
Dist. Columbia	93	2,914	66	148	255	2,577	2,386	13	2,399
Florida	95	3,267	202	214	273	2,982	2,818	10	2,828
Georgia	95	3,117	138	202	154	2,899	2,759	7	2,766
Hawaii	95	3,009	89	168	169	2,761	2,617	8	2,625
Idaho	97	2,983	90	100	102	2,871	2,777	7	2,784
Indiana	96	2,815	72	86	92	2,709	2,590	3	2,593
Iowa	96	3,001	54	74	98	2,883	2,759	11	2,770
Kentucky	96	2,970	109	156	99	2,824	2,690	13	2,703
Louisiana	95	3,113	102	155	122	2,938	2,776	16	2,792
Maine[3]	95	2,161	31	46	124	2,022	1,920	3	1,923
Maryland	96	3,170	103	175	126	2,972	2,842	2	2,844
Massachusetts	95	2,942	32	77	219	2,678	2,544	5	2,549
Michigan[4]	94	2,736	82	100	136	2,582	2,417	6	2,423
Minnesota[3]	95	2,924	39	60	104	2,799	2,666	3	2,669
Mississippi	97	3,023	89	159	146	2,807	2,709	3	2,712
Missouri	96	2,778	112	152	117	2,621	2,501	8	2,509
Nebraska	96	2,602	44	80	122	2,444	2,320	17	2,337
New Hampshire[2]	96	2,538	47	78	99	2,408	2,309	7	2,316
New Jersey	96	2,483	49	77	133	2,322	2,220	11	2,231
New Mexico[3]	95	2,874	50	184	188	2,552	2,438	0	2,438
New York	96	2,545	44	75	127	2,387	2,277	7	2,284
North Carolina	95	3,144	142	142	122	3,022	2,880	4	2,884
North Dakota	96	2,312	42	40	45	2,269	2,190	3	2,193
Ohio	95	2,962	84	113	166	2,767	2,632	5	2,637
Oklahoma[1]	84	2,936	110	149	215	2,682	2,250	4	2,254
Pennsylvania	96	3,015	55	90	112	2,868	2,729	11	2,740
Rhode Island	95	2,767	54	142	161	2,518	2,390	0	2,390
South Carolina	97	3,045	110	143	144	2,868	2,771	0	2,771
Tennessee	96	2,979	107	148	117	2,821	2,704	4	2,708
Texas	96	3,013	105	162	234	2,722	2,618	5	2,623
Utah	96	3,130	85	167	128	2,930	2,793	6	2,799
Virginia	95	3,105	130	146	163	2,926	2,777	9	2,786
West Virginia	96	3,068	72	92	134	2,914	2,782	4	2,786
Wisconsin	96	3,079	61	89	141	2,910	2,793	4	2,797
Wyoming	96	2,833	98	118	98	2,717	2,602	3	2,605
TERRITORIES									
Guam	95	2,158	104	91	133	2,038	1,914	19	1,933
Virgin Islands	97	952	22	18	24	932	905	0	905

See explanations of the notations and guidelines about sample representativeness and for the derivation of weighted participation. Weighted percentages for the nation and region are based on schools sampled for all subject areas assessed in 1990 (reading, science, and mathematics) or 1992 (mathematics, reading, and writing). However, based on the national sampling design, the rates shown also are the best estimates for the mathematics assessment. [1]The weighted student response rate within participating schools was below 85 percent. Oklahoma, however, was the only state that required parental permission forms on a statwide basis. [2]The nonresponding students within participating schools included a class of students with similar characteristics, who together comprised more than five percent of the state's weighted assessable student sample. Student groups from which a state needed minimum levels of participation were determined by age of student and type of assessment session (unmonitored or monitored), as well as school urbanicity, minority enrollment, and median household income of the area in which the school is located.

BEST COPY AVAILABLE

TABLE B.2 | Student Participation Rates (continued)

PUBLIC SCHOOLS	Grade 8 - 1992 Weighted Percentage Student Participation After Make-ups	Number Students Original Sample	Number Students Supplemental Sample	Number Students Withdrawn	Number Students Excluded	Number Students to be Assessed	Number Students Assessed Initial Sessions	Number Students Assessed Make-ups	Total Number Students Assessed
NATION	89	7,406	–	–	582	6,824	5,975	58	6,033
Northeast	89	1,321	–	–	100	1,221	1,041	20	1,061
Southeast	90	1,885	–	–	92	1,793	1,807	3	1,810
Central	89	1,872	–	–	103	1,769	1,392	6	1,398
West	88	2,528	–	–	287	2,24[illegible]	1,935	29	1,964
STATES									
Alabama[3]	95	3,011	65	183	165	2,748	2,611	12	2,623
Arizona	93	3,089	181	280	178	2,812	2,565	52	2,617
Arkansas	94	2,978	84	169	176	2,717	2,540	16	2,556
California[3]	92	3,101	120	212	246	2,763	2,510	27	2,537
Colorado	93	3,199	126	183	136	3,006	2,773	26	2,799
Connecticut	94	3,029	71	125	192	2,783	2,590	23	2,613
Delaware	92	2,220	83	106	97	2,098	1,858	76	1,934
Dist. Columbia	85	2,517	79	234	225	2,137	1,692	124	1,816
Florida	91	3,073	184	246	199	2,812	2,515	34	2,549
Georgia	93	3,011	133	220	137	2,787	2,563	26	2,589
Hawaii	90	2,904	85	123	142	2,724	2,421	33	2,454
Idaho	95	2,936	79	125	91	2,799	2,638	7	2,645
Indiana	94	3,000	49	89	140	2,820	2,645	14	2,659
Iowa	95	3,133	40	85	129	2,959	2,801	15	2,816
Kentucky	96	3,087	87	156	135	2,883	2,746	10	2,756
Louisiana	92	3,028	80	194	120	2,794	2,565	17	2,582
Maine[2]	93	2,838	32	48	124	2,698	2,512	8	2,520
Maryland	92	2,803	108	178	128	2,605	2,364	35	2,399
Massachusetts	94	2,909	24	93	217	2,623	2,439	17	2,456
Michigan	94	3,020	79	122	184	2,793	2,573	43	2,616
Minnesota	94	2,758	38	85	92	2,619	2,439	32	2,471
Mississippi	95	2,958	76	191	207	2,636	2,490	6	2,496
Missouri	95	2,984	124	165	128	2,815	2,641	25	2,666
Nebraska	96	2,543	31	74	108	2,392	2,283	2	2,285
New Hampshire[3]	94	2,958	49	96	156	2,755	2,562	20	2,582
New Jersey	94	2,506	50	80	169	2,307	2,160	14	2,174
New Mexico[3]	93	3,041	70	168	183	2,780	2,556	29	2,585
New York	92	2,581	44	85	193	2,347	2,131	27	2,158
North Carolina	94	3,071	114	147	102	2,936	2,759	10	2,769
North Dakota	96	2,513	33	65	63	2,418	2,305	9	2,314
Ohio	93	2,942	87	120	177	2,732	2,518	17	2,535
Oklahoma[4]	80	2,934	114	154	184	2,710	2,129	12	2,141
Pennsylvania[3]	94	2,964	32	63	127	2,806	2,611	29	2,640
Rhode Island	93	2,481	45	118	119	2,289	2,099	21	2,120
South Carolina	94	3,057	103	178	174	2,808	2,822	3	2,825
Tennessee	94	2,838	117	174	137	2,644	2,470	15	2,485
Texas	94	3,048	133	182	205	2,794	2,596	18	2,614
Utah	94	3,124	102	175	141	2,910	2,713	13	2,726
Virginia	94	3,091	103	169	153	2,872	2,690	20	2,710
West Virginia	94	3,097	43	119	178	2,843	2,675	15	2,690
Wisconsin	94	3,165	58	91	130	3,002	2,787	27	2,814
Wyoming	95	2,743	64	124	107	2,576	2,403	41	2,444
TERRITORIES									
Guam	90	1,734	56	51	72	1,667	1,491	5	1,496
Virgin Islands	92	1,708	39	60	86	1,601	1,410	69	1,479

[3] One or more schools in the original sample initially declined and then decided to participate after their substitute(s) had also agreed to participate. Further, assessments were conducted in both the original and substitute schools. For these cases, the students in the substitute school(s) are included in the counts of students in the table. The state's estimates will be based on the student responses from the original school only. [4] In one or more schools an assessment was conducted but the wrong materials were sent to the school(s). The students in these school(s) are included in the counts of students in the tables. However, the state's estimates will not be based on these student responses.

TABLE B.2 | Student Participation Rates (continued)

PUBLIC SCHOOLS	Grade 8 - 1990								
	Weighted Percentage Student Participation After Make-ups	Number Students Original Sample	Number Students Supplemental Sample	Number Students Withdrawn	Number Students Excluded	Number Students to be Assessed	Number Students Assessed Initial Sessions	Number Students Assessed Make-ups	Total Number Students Assessed
NATION	90	11,871	–	–	741	11,130	9,775	147	9,922
Northeast	91	1,922	–	–	96	1,826	1,822	11	1,833
Southeast	91	3,163	–	–	119	3,044	2,752	0	2,752
Central	91	2,481	–	–	218	2,272	2,017	22	2,039
West	88	4,295	–	–	307	3,988	3,384	114	3,498
STATES									
Alabama	95	2,908	99	186	162	2,659	2,511	20	2,531
Arizona	93	2,945	161	206	158	2,742	2,480	78	2,558
Arkansas	95	3,104	127	183	244	2,804	2,640	29	2,669
California	93	2,933	63	135	242	2,619	2,353	71	2,424
Colorado	94	3,074	103	192	142	2,843	2,632	43	2,675
Connecticut	95	3,085	58	115	213	2,815	2,846	26	2,872
Delaware	93	2,455	83	163	122	2,253	2,052	58	2,110
Dist. Columbia	88	2,758	72	237	156	2,437	2,017	118	2,135
Florida	92	3,005	148	209	200	2,744	2,475	59	2,534
Georgia	94	3,175	126	254	117	2,930	2,736	30	2,766
Hawaii	93	2,933	82	120	151	2,744	2,452	99	2,551
Idaho	96	2,941	90	123	78	2,830	2,707	9	2,716
Indiana	95	2,910	81	143	144	2,704	2,534	35	2,569
Iowa	96	2,714	40	73	104	2,577	2,462	12	2,474
Kentucky	95	3,068	88	179	158	2,819	2,660	20	2,680
Louisiana	94	2,949	108	204	130	2,723	2,544	28	2,572
Maine	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
Maryland	94	3,151	82	115	152	2,966	2,732	62	2,794
Massachusetts	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
Michigan	95	2,941	64	140	129	2,736	2,524	63	2,587
Minnesota	95	2,857	50	105	87	2,715	2,537	47	2,584
Mississippi	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
Missouri	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
Nebraska	95	2,766	58	93	84	2,647	2,497	22	2,519
New Hampshire	95	2,870	52	80	132	2,710	2,548	20	2,568
New Jersey	94	3,149	63	113	234	2,865	2,675	35	2,710
New Mexico	94	3,091	122	236	185	2,792	2,600	43	2,643
New York	93	2,704	56	98	171	2,491	2,242	60	2,302
North Carolina	95	3,180	87	142	107	3,008	2,791	52	2,843
North Dakota	96	2,672	55	58	91	2,578	2,483	2	2,485
Ohio	95	3,030	90	138	174	2,808	2,842	31	2,873
Oklahoma*	80	3,007	107	194	164	2,756	2,208	14	2,222
Pennsylvania	94	2,848	51	77	148	2,675	2,506	22	2,528
Rhode Island	93	3,152	91	178	208	2,857	2,633	42	2,675
South Carolina	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
Tennessee	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
Texas	96	2,909	140	196	196	2,657	2,525	17	2,542
Utah	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
Virginia	94	3,120	85	195	174	2,836	2,833	28	2,861
West Virginia	94	3,008	77	152	172	2,761	2,532	68	2,600
Wisconsin	94	3,101	52	92	145	2,916	2,705	45	2,750
Wyoming	96	2,973	83	126	106	2,824	2,662	39	2,701
TERRITORIES									
Guam	93	1,810	62	58	75	1,739	1,573	44	1,617
Virgin Islands	93	1,490	1	16	48	1,427	1,299	27	1,326

(–) Because student sampling for the national assessment was implemented within several days of the assessment within each school there was no supplemental sample and the number of students withdrawn was negligible. (xxx) Did not participate in the 1990 Trial State Assessment.

TABLE B.3 | Summary of School and Student Participation

PUBLIC SCHOOLS	Grade 4 - 1992: Weighted Percentage School Participation Before Substitution	Weighted Percentage School Participation After Substitution	Notation Number 1	Weighted Percentage Student Participation After Make-ups	Notation Number 3	Weighted Overall Rate
NATION	86	86		94		81
Northeast	82	82		94		78
Southeast	94	94		93		88
Central	92	92		94		87
West	79	79		94		75
STATES						
Alabama	75	97		95		93
Arizona	100	100		95		95
Arkansas	90	99		96		95
California	91	97		94		91
Colorado	100	100		95		95
Connecticut	99	99		96		95
Delaware	92	92		95		87
Dist. Columbia	99	99		93		92
Florida	100	100		95		95
Georgia	100	100		95		95
Hawaii	100	100		95		95
Idaho	84	97		97		94
Indiana	76	91		96		87
Iowa	100	100		96		96
Kentucky	93	96		96		92
Louisiana	100	100		95		95
Maine	57	71	***	95		68
Maryland	99	99		96		95
Massachusetts	87	97		95		92
Michigan	83	90		94		84
Minnesota	82	94		95		89
Mississippi	98	100		97		97
Missouri	89	97		96		93
Nebraska	80	87	***	96		83
New Hampshire	69	80	***	96		77
New Jersey	76	82	***	96		79
New Mexico	75	90		95		86
New York	78	83	***	96		80
North Carolina	95	99		95		94
North Dakota	73	90		96		87
Ohio	79	91		95		87
Oklahoma	86	98		84	***	83
Pennsylvania	84	95		96		91
Rhode Island	83	96		95		91
South Carolina	98	99		97		96
Tennessee	92	93		96		89
Texas	93	98		96		94
Utah	99	99		96		95
Virginia	99	99		95		94
West Virginia	100	100		96		96
Wisconsin	100	100		96		96
Wyoming	97	97		96		93
TERRITORIES						
Guam	94	94		95		89
Virgin Islands	100	100		97		97

See explanations of the notations and guidelines about sample representativeness and for the derivation of weighted participation. Weighted percentages for the nation and region are based on schools sampled for all subject areas assessed in 1990 (reading, science, and mathematics) or 1992 (mathematics, reading, and writing). However, based on the national sampling design, the rates shown also are the best estimates for the mathematics assessment. Notation Number 1 = Both the state's weighted participation rate for the initial sample of schools was below 85% AND the weighted school participation rate after substitution was below 90%; OR the weighted school participation rate of the initial sample of schools was below 70% (regardless of the participation rate after substitution.) Notation number 3 = The weighted student response rate within participating schools was below 85 percent.

TABLE B.3 | Summary of School and Student Participation (continued)

PUBLIC SCHOOLS	Grade 8 - 1992						Grade 8 - 1990					
	Weighted Percentage School Participation Before Substitution	Weighted Percentage School Participation After Substitution	Notation Number 1	Weighted Percentage Student Participation After Make-ups	Notation Number 3	Weighted Overall Rate	Weighted Percentage School Participation Before Substitution	Weighted Percentage School Participation After Substitution	Notation Number 1	Weighted Percentage Student Participation After Make-ups	Notation Number 3	Weighted Overall Rate
NATION	88	89		89		79	88	92		90		83
Northeast	92	92		89		82	72	90		91		82
Southeast	94	94		90		85	94	94		91		86
Central	86	87		89		78	94	94		91		86
West	82	84		88		74	88	90		88		79
STATES												
Alabama	66	92	***	95		88	85	97		95		93
Arizona	99	99		93		92	97	97		93		90
Arkansas	89	97		94		91	100	100		95		95
California	93	98		92		90	94	94		93		87
Colorado	100	100		93		93	100	100		94		94
Connecticut	99	99		94		93	100	100		95		95
Delaware	100	100		92		92	100	100		93		93
Dist. Columbia	100	100		85		85	100	100		88		88
Florida	100	100		91		91	98	98		92		90
Georgia	99	99		93		92	100	100		94		94
Hawaii	100	100		90		90	100	100		93		93
Idaho	85	91		95		86	97	97		96		93
Indiana	79	94		94		88	89	94		95		89
Iowa	99	99		95		94	91	91		96		88
Kentucky	96	98		96		94	100	100		95		95
Louisiana	100	100		92		92	100	100		94		94
Maine	62	84	***	93		78	xxx	xxx		xxx		xxx
Maryland	89	91		92		84	100	100		94		94
Massachusetts	83	95		94		89	xxx	xxx		xxx		xxx
Michigan	78	94		94		88	90	97		95		92
Minnesota	81	92		94		87	90	93		95		89
Mississippi	99	100		95		95	xxx	xxx		xxx		xxx
Missouri	92	99		95		94	xxx	xxx		xxx		xxx
Nebraska	75	85	***	96		81	87	94		95		90
New Hampshire	80	92		94		86	91	97		95		92
New Jersey	69	78	***	94		73	97	98		94		93
New Mexico	77	94		93		87	100	100		94		94
New York	81	83	***	92		77	86	86		93		79
North Carolina	94	98		94		92	100	100		95		95
North Dakota	78	97		96		93	96	100		96		96
Ohio	77	90		93		83	96	98		95		93
Oklahoma	82	98		80	***	79	78	99		80	***	79
Pennsylvania	81	94		94		89	90	93		94		88
Rhode Island	85	100		93		92	94	97		93		91
South Carolina	94	97		94		91	xxx	xxx		xxx		xxx
Tennessee	87	91		94		86	xxx	xxx		xxx		xxx
Texas	95	99		94		93	88	97		96		93
Utah	100	100		94		94	xxx	xxx		xxx		xxx
Virginia	97	97		94		92	99	99		94		93
West Virginia	100	100		94		94	100	100		94		94
Wisconsin	100	100		94		94	99	99		94		93
Wyoming	99	99		95		94	100	100		96		96
TERRITORIES												
Guam	100	100		90		90	100	100		93		93
Virgin Islands	100	100		92		92	100	100		93		93

(xxx) Did not participate in the 1990 Trial State Assessment.

TABLE B.4 | Weighted Percentages of Students Excluded (IEP and LEP) from Original Sample

PUBLIC SCHOOLS	Grade 4 - 1992					
	Total Percentage Students Identified IEP and LEP	Total Percentage Students Excluded	Percentage Students Identified IEP	Percentage Students Excluded · IEP	Percentage Students Identified LEP	Percentage Students Excluded LEP
NATION	12	8	9	6	4	3
Northeast	12	8	9	5	3	3
Southeast	11	7	9	6	1	1
Central	7	5	6	4	1	1
West	16	12	10	6	9	7
STATES						
Alabama	10	5	10	4	0	0
Arizona	15	5	7	3	9	2
Arkansas	12	5	11	5	1	0
California	28	12	8	3	22	10
Colorado	10	5	8	4	2	1
Connecticut	14	7	10	4	4	3
Delaware	12	5	11	5	1	1
Dist. Columbia	12	9	8	7	4	2
Florida	17	8	13	7	4	2
Georgia	10	5	9	5	1	1
Hawaii	14	6	10	5	4	2
Idaho	9	3	8	3	2	1
Indiana	7	3	6	3	1	0
Iowa	9	3	8	3	1	0
Kentucky	8	3	8	3	0	0
Louisiana	8	4	7	4	1	0
Maine	14	6	14	6	0	0
Maryland	11	4	10	3	1	1
Massachusetts	18	7	15	6	3	2
Michigan	7	5	7	5	1	1
Minnesota	9	3	7	3	2	0
Mississippi	7	5	7	5	0	0
Missouri	12	4	12	4	0	0
Nebraska	13	4	12	4	1	0
New Hampshire	12	4	12	4	0	0
New Jersey	11	6	8	3	4	2
New Mexico	15	7	12	6	3	1
New York	12	5	7	3	5	--2
North Carolina	12	4	12	3	1	0
North Dakota	9	2	8	2	1	0
Ohio	10	6	10	6	1	0
Oklahoma	14	7	12	7	2	0
Pennsylvania	9	4	8	3	1	1
Rhode Island	16	6	10	4	6	3
South Carolina	10	5	10	5	0	0
Tennessee	11	4	11	4	0	0
Texas	17	8	9	5	9	4
Utah	10	4	9	4	1	1
Virginia	12	5	10	5	1	1
West Virginia	9	4	9	4	0	0
Wisconsin	11	5	9	5	2	1
Wyoming	10	4	9	3	1	0
TERRITORIES						
Guam	12	6	8	4	7	3
Virgin Islands	5	3	3	2	3	1

IEP = Individual Education Plan and LEP = Limited English Proficiency. To be excluded, a student was supposed to be IEP or LEP and judged incapable of participating in the assessment. A student reported as both IEP and LEP is counted once in the overall rate (first column), once in the overall excluded rate (second column), and separately in the remaining columns. Note: Weighted percentages for the nation and region are based on students sampled for all subject areas assessed in 1990 (reading, science, and mathematics) or 1992 (mathematics, reading, and writing). However, based on the national sampling design, the rates shown also are the best estimates for the mathematics assessment.

TABLE B.4 | Weighted Percentages of Students Excluded (IEP and LEP) from Original Sample (continued)

PUBLIC SCHOOLS	Grade 8 - 1992						Grade 8 - 1990					
	Total Percentage Students Identified IEP and LEP	Total Percentage Students Excluded	Percentage Students Identified IEP	Percentage Students Excluded IEP	Percentage Students Identified LEP	Percentage Students Excluded LEP	Total Percentage Students Identified IEP and LEP	Total Percentage Students Excluded	Percentage Students Identified IEP	Percentage Students Excluded IEP	Percentage Students Identified LEP	Percentage Students Excluded LEP
NATION	12	7	9	5	3	2	8	6	6	5	2	1
Northeast	12	8	10	6	3	2	6	4	6	4	1	0
Southeast	11	6	10	5	1	1	6	4	6	4	0	0
Central	9	6	8	5	1	1	9	8	8	6	2	2
West	15	9	8	5	8	4	10	6	6	4	4	2
STATES												
Alabama	10	5	10	5	0	0	10	6	10	6	0	0
Arizona	12	6	6	4	6	2	13	5	7	4	6	2
Arkansas	11	6	11	6	0	0	12	8	11	8	0	0
California	20	8	8	4	13	5	16	8	7	4	9	5
Colorado	9	4	8	4	1	1	10	5	9	4	1	1
Connecticut	14	7	12	5	3	1	12	7	10	6	2	1
Delaware	10	4	9	4	1	0	10	5	9	4	1	1
Dist. Columbia	12	10	9	7	3	2	7	6	5	5	1	1
Florida	13	6	10	5	4	2	12	7	9	5	3	2
Georgia	8	5	7	4	1	0	7	4	7	4	0	0
Hawaii	13	5	9	3	4	2	10	5	7	4	3	1
Idaho	7	3	6	3	1	0	7	3	6	2	1	0
Indiana	9	5	8	4	1	0	8	5	7	5	0	0
Iowa	11	4	10	4	1	0	10	4	10	4	0	0
Kentucky	9	5	9	5	0	0	8	5	8	5	0	0
Louisiana	7	4	7	4	0	0	7	5	6	4	0	0
Maine	11	4	11	4	0	0	xxx	xxx	xxx	xxx	xxx	xxx
Maryland	11	5	10	4	1	1	11	5	10	4	1	1
Massachusetts	18	8	15	6	4	2	xxx	xxx	xxx	xxx	xxx	xxx
Michigan	9	6	9	6	1	0	9	5	8	4	1	0
Minnesota	8	3	7	3	1	0	9	3	8	3	1	0
Mississippi	10	7	10	7	0	0	xxx	xxx	xxx	xxx	xxx	xxx
Missouri	11	4	10	4	0	0	xxx	xxx	xxx	xxx	xxx	xxx
Nebraska	11	4	10	4	1	0	9	3	8	3	0	0
New Hampshire	13	5	12	5	0	0	12	5	12	5	0	0
New Jersey	14	7	12	6	3	1	13	8	10	6	2	2
New Mexico	12	5	11	4	3	1	10	7	9	6	2	1
New York	13	8	10	6	3	3	12	7	9	5	4	2
North Carolina	12	3	12	3	1	0	9	3	9	3	0	0
North Dakota	8	2	7	2	1	0	8	3	8	3	1	0
Ohio	10	6	10	6	0	0	8	6	8	6	0	0
Oklahoma	11	6	10	6	1	0	9	6	8	5	1	0
Pennsylvania	10	4	9	4	1	0	11	6	10	5	1	0
Rhode Island	14	5	10	4	4	2	15	7	12	5	4	2
South Carolina	10	6	10	6	0	0	xxx	xxx	xxx	xxx	xxx	xxx
Tennessee	10	5	10	5	0	0	xxx	xxx	xxx	xxx	xxx	xxx
Texas	14	7	9	5	6	2	14	7	8	5	5	2
Utah	9	4	9	4	1	0	xxx	xxx	xxx	xxx	xxx	xxx
Virginia	12	5	10	5	2	1	10	6	8	5	2	1
West Virginia	10	6	10	6	0	0	10	6	10	6	0	0
Wisconsin	10	4	10	4	1	0	8	5	8	4	1	0
Wyoming	9	4	9	4	0	0	9	4	8	4	1	0
TERRITORIES												
Guam	7	4	5	3	2	1	7	4	5	4	2	1
Virgin Islands	7	5	5	3	2	2	4	3	4	3	0	0

(xxx) Did not participate in the 1990 Trial State Assessment.

BEST COPY AVAILABLE

TABLE B.5 | Weighted Percentages of Absent, IEP, and LEP Students Based on Those Invited to Participate in the Assessment

PUBLIC SCHOOLS	Grade 4 - 1992					
	Weighted Percentage Student Participation After Make-Ups	Weighted Percentage Absent	Weighted Percentage Assessed IEP	Weighted Percentage Absent IEP	Weighted Percentage Assessed LEP	Weighted Percentage Absent LEP
NATION	94	6	89	11	93	7
Northeast	94	6	93	7	81	19
Southeast	93	7	83	17	68	32
Central	94	6	92	8	96	4
West	93	7	90	10	94	6
STATES						
Alabama	95	5	95	5	100	0
Arizona	95	5	96	4	93	7
Arkansas	96	4	95	5	100	0
California	94	6	85	15	93	7
Colorado	95	5	92	8	95	5
Connecticut	96	4	94	6	95	5
Delaware	95	5	95	5	65	35
Dist. Columbia	93	7	85	15	82	18
Florida	95	5	95	5	95	5
Georgia	95	5	92	8	90	10
Hawaii	95	5	91	9	95	5
Idaho	97	3	93	7	95	5
Indiana	96	4	94	6	86	14
Iowa	96	4	96	4	100	0
Kentucky	96	4	90	10	100	0
Louisiana	95	5	91	9	100	0
Maine	95	5	95	5	100	0
Maryland	96	4	95	5	100	0
Massachusetts	95	5	93	7	92	8
Michigan	94	6	93	7	100	0
Minnesota	95	5	92	8	95	5
Mississippi	97	3	95	5	100	0
Missouri	96	4	97	3	100	0
Nebraska	96	4	94	6	92	8
New Hampshire	96	4	94	6	91	9
New Jersey	96	4	95	5	100	0
New Mexico	95	5	93	7	97	3
New York	96	4	91	9	97	3
North Carolina	95	5	91	9	100	0
North Dakota	96	4	96	4	100	0
Ohio	95	5	95	5	100	0
Oklahoma	85	15	72	28	69	31
Pennsylvania	96	4	99	1	100	0
Rhode Island	95	5	94	6	94	6
South Carolina	97	3	91	9	100	0
Tennessee	96	4	96	4	100	0
Texas	96	4	98	2	98	2
Utah	96	4	95	5	100	0
Virginia	95	5	90	10	95	5
West Virginia	96	4	94	6	0	0
Wisconsin	96	4	91	9	93	7
Wyoming	96	4	94	6	89	11
TERRITORIES						
Guam	95	5	85	15	88	12
Virgin Islands	97	3	66	34	95	5

IEP = Individual Education Plan and LEP = Limited English Proficiency. Note: Weighted percentages for the nation and region are based on students sampled for all subject areas assessed in 1990 (reading, science, and mathematics) or 1992 (mathematics, reading, and writing). However, based on the national sampling design, the rates shown also are the best estimates for the mathematics assessment.

TABLE B.5 | Weighted Percentages of Absent, IEP, and LEP Students Based on Those Invited to Participate in the Assessment (continued)

PUBLIC SCHOOLS	Grade 8 - 1992						Grade 8 - 1990					
	Weighted Percentage Student Participation After Make-Ups	Weighted Percentage Absent	Weighted Percentage Assessed IEP	Weighted Percentage Absent IEP	Weighted Percentage Assessed LEP	Weighted Percentage Absent LEP	Weighted Percentage Student Participation After Make-Ups	Weighted Percentage Absent	Weighted Percentage Assessed IEP	Weighted Percentage Absent IEP	Weighted Percentage Assessed LEP	Weighted Percentage Absent LEP
NATION	89	11	80	20	84	16	90	10	91	9	87	13
Northeast	89	11	78	22	78	22	91	9	86	14	81	19
Southeast	89	11	78	22	78	22	91	9	93	7	0	0
Central	91	9	84	16	84	16	91	9	98	2	100	0
West	88	12	83	17	86	14	88	12	88	12	87	13
STATES												
Alabama	95	5	90	10	74	26	95	5	92	8	100	0
Arizona	93	7	90	10	97	3	93	7	90	10	89	11
Arkansas	94	6	91	9	100	0	95	5	91	9	100	0
California	92	8	87	13	90	10	93	7	97	3	94	6
Colorado	93	7	92	8	100	0	94	6	92	8	100	0
Connecticut	94	6	90	10	100	0	95	5	93	7	100	0
Delaware	92	8	86	14	100	0	93	7	94	6	80	20
Dist. Columbia	85	15	63	37	89	11	88	12	92	8	0	0
Florida	91	9	82	18	94	6	92	8	88	12	79	21
Georgia	93	7	86	14	92	8	94	6	97	3	93	7
Hawaii	90	10	81	19	93	7	93	7	85	15	100	0
Idaho	95	5	95	5	100	0	96	4	97	3	100	0
Indiana	94	6	92	8	100	0	95	5	93	7	100	0
Iowa	95	5	91	9	100	0	96	4	97	3	100	0
Kentucky	96	4	97	3	100	0	95	5	94	6	100	0
Louisiana	92	8	87	13	79	21	94	6	96	4	100	0
Maine	93	7	87	13	100	0	xxx	xxx	xxx	xxx	xxx	xxx
Maryland	92	8	86	14	100	0	94	6	88	12	100	0
Massachusetts	93	7	86	14	92	8	xxx	xxx	xxx	xxx	xxx	xxx
Michigan	94	6	96	4	100	0	95	5	94	6	100	0
Minnesota	94	6	88	12	72	28	95	5	96	4	100	0
Mississippi	95	5	91	9	100	0	xxx	xxx	xxx	xxx	xxx	xxx
Missouri	95	5	95	5	100	0	xxx	xxx	xxx	xxx	xxx	xxx
Nebraska	95	5	92	8	100	0	95	5	95	5	100	0
New Hampshire	94	6	89	11	100	0	95	5	95	5	100	0
New Jersey	94	6	92	8	97	3	94	6	88	12	94	6
New Mexico	93	7	88	12	89	11	94	6	95	5	95	5
New York	91	9	87	13	100	0	93	7	94	6	100	0
North Carolina	94	6	93	7	78	22	95	5	93	7	100	0
North Dakota	96	4	95	5	100	0	96	4	95	5	100	0
Ohio	93	7	86	14	91	9	95	5	97	3	100	0
Oklahoma	80	20	68	32	73	27	80	20	76	24	100	0
Pennsylvania	94	6	87	13	83	17	94	6	95	5	100	0
Rhode Island	93	7	90	10	91	9	93	7	92	8	91	9
South Carolina	93	7	88	12	100	0	xxx	xxx	xxx	xxx	xxx	xxx
Tennessee	94	6	91	9	100	0	xxx	xxx	xxx	xxx	xxx	xxx
Texas	94	6	93	7	85	15	96	4	97	3	94	6
Utah	94	6	90	10	100	0	xxx	xxx	xxx	xxx	xxx	xxx
Virginia	94	6	90	10	96	4	94	6	91	9	90	10
West Virginia	94	6	92	8	100	0	94	6	94	6	100	0
Wisconsin	94	6	87	13	86	14	94	6	93	7	93	7
Wyoming	95	5	94	6	57	43	96	4	93	7	100	0
TERRITORIES												
Guam	90	10	86	14	83	17	93	7	75	25	100	0
Virgin Islands	92	8	53	47	89	11	93	7	73	27	100	0

(xxx) Did not participate in the 1990 Trial State Assessment.

TABLE B.6 | Questionnaire Response Rates

PUBLIC SCHOOLS	Grade 4 - 1992				
	Weighted Percentage of Students Matched to Mathematics Teacher Questionnaires	Percentage of Mathematics Teacher Questionnaires Returned	Weighted Percentage of Students Matched to School Characteristics/ Policies Questionnaires	Percentage of School Characteristics/ Policies Questionnaires Returned	Percentage of Excluded Student Questionnaires Returned
NATION	74.9	97.7	99.0	98.5	91.0
Northeast	75.9	95.8	100.0	100.0	94.6
Southeast	82.7	99.0	96.3	96.1	94.4
Central	72.8	97.6	99.8	98.4	93.3
West	67.8	97.2	100.0	100.0	87.1
STATES					
Alabama	91.0	100.0	100.0	100.0	96.1
Arizona	92.5	99.6	99.0	99.1	96.0
Arkansas	94.3	100.0	100.0	100.0	100.0
California	90.6	99.3	98.9	99.1	89.8
Colorado	86.3	99.3	100.0	100.0	95.2
Connecticut	89.0	99.8	98.6	98.2	88.3
Delaware	94.3	100.0	100.0	100.0	99.2
Dist. Columbia	80.0	99.0	93.7	94.4	94.1
Florida	90.1	98.9	99.2	99.1	97.1
Georgia	88.2	99.3	100.0	100.0	97.4
Hawaii	92.5	98.8	98.8	99.1	98.2
Idaho	91.8	99.7	100.0	100.0	98.0
Indiana	89.3	100.0	100.0	100.0	98.9
Iowa	90.8	99.5	100.0	100.0	98.0
Kentucky	89.9	99.5	99.4	99.1	100.0
Louisiana	91.2	99.6	98.9	99.1	96.7
Maine	90.6	99.1	99.3	98.9	94.3
Maryland	90.2	99.5	100.0	100.0	92.9
Massachusetts	87.4	100.0	100.0	100.0	97.7
Michigan	89.9	100.0	100.0	100.0	95.6
Minnesota	80.4	97.6	95.5	96.2	91.2
Mississippi	88.6	99.8	100.0	100.0	97.3
Missouri	90.6	99.7	100.0	100.0	91.5
Nebraska	86.1	100.0	99.0	99.2	98.2
New Hampshire	94.9	99.7	97.8	99.0	100.0
New Jersey	92.5	100.0	100.0	100.0	99.2
New Mexico	87.3	99.0	100.0	100.0	94.8
New York	91.7	99.0	99.5	98.9	100.0
North Carolina	94.4	100.0	99.1	99.1	99.2
North Dakota	93.1	100.0	100.0	100.0	97.8
Ohio	89.6	99.5	99.8	99.1	94.0
Oklahoma	94.5	99.1	97.9	98.4	92.6
Pennsylvania	93.8	100.0	100.0	100.0	100.0
Rhode Island	92.1	99.4	99.1	98.9	93.8
South Carolina	97.3	99.6	100.0	100.0	98.6
Tennessee	93.4	100.0	99.3	98.1	97.4
Texas	84.5	99.9	99.3	99.0	99.1
Utah	95.2	99.5	100.0	100.0	97.7
Virginia	88.4	99.6	97.7	97.3	95.7
West Virginia	88.5	100.0	100.0	100.0	100.0
Wisconsin	90.2	99.7	99.3	99.2	97.1
Wyoming	91.4	100.0	99.9	99.3	100.0
TERRITORIES					
Guam	97.5	98.3	93.7	95.0	87.2
Virgin Islands	83.4	97.7	100.0	100.0	100.0

The Mathematics Teacher Questionnaire requested background information about the teacher (Part I) and information about instruction in particular classes (Part II). The percentage of students matched to questionnaires is provided for Part II. If they differed, the match rates for Part I were higher. Note: For the nation and regions, the percentage of excluded student questionnaires returned is based on students sampled for all subjects assessed in 1990 (reading, science, and mathematics) or 1992 (mathematics, reading, and writing). However, based on the sampling design, these rates also are the best estimates of the comparable rates for the mathematics assessment in each year.

TABLE B.6 | Questionnaire Response Rates (continued)

PUBLIC SCHOOLS	Grade 8 - 1992				
	Weighted Percentage of Students Matched to Mathematics Teacher Questionnaires	Percentage of Mathematics Teacher Questionnaires Returned	Weighted Percentage of Students Matched to School Characteristics/ Policies Questionnaires	Percentage of School Characteristics/ Policies Questionnaires Returned	Percentage of Excluded Student Questionnaires Returned
NATION	79.7	96.2	94.5	92.7	88.8
Northeast	81.7	97.9	89.0	90.0	88.6
Southeast	80.2	97.9	98.5	96.2	85.1
Central	78.1	95.1	95.0	92.8	94.5
West	80.8	94.8	94.3	89.7	88.0
STATES					
Alabama	94.1	100.0	100.0	100.0	99.4
Arizona	91.1	98.1	97.7	98.0	97.2
Arkansas	95.0	99.6	99.0	99.0	100.0
California	92.5	98.2	100.0	100.0	97.9
Colorado	90.5	99.0	98.2	98.2	98.5
Connecticut	97.3	98.7	99.4	99.0	96.9
Delaware	97.4	100.0	100.0	100.0	97.9
Dist. Columbia	83.4	96.8	93.6	97.1	92.0
Florida	94.8	99.0	98.2	99.0	97.0
Georgia	94.1	99.7	98.2	98.0	98.5
Hawaii	89.8	96.4	97.9	98.0	97.9
Idaho	90.8	98.6	96.2	97.1	97.8
Indiana	92.4	100.0	100.0	100.0	96.4
Iowa	92.4	99.6	100.0	100.0	100.0
Kentucky	91.0	99.3	96.2	96.2	99.3
Louisiana	93.8	99.7	100.0	100.0	100.0
Maine	91.4	100.0	100.0	100.0	95.9
Maryland	90.9	99.3	98.2	97.8	93.0
Massachusetts	89.7	100.0	96.5	97.9	98.2
Michigan	93.2	99.7	100.0	100.0	99.5
Minnesota	83.8	98.0	95.7	96.7	95.7
Mississippi	95.1	99.6	99.0	99.0	98.1
Missouri	95.9	99.7	100.0	100.0	98.4
Nebraska	93.8	100.0	94.4	95.2	100.0
New Hampshire	92.9	98.8	100.0	100.0	98.1
New Jersey	96.2	100.0	100.0	100.0	99.4
New Mexico	92.9	99.3	100.0	100.0	97.5
New York	94.6	100.0	100.0	100.0	100.0
North Carolina	95.3	100.0	100.0	100.0	100.0
North Dakota	96.9	100.0	100.0	100.0	100.0
Ohio	89.6	99.6	97.4	96.9	98.3
Oklahoma	91.7	99.3	99.3	99.0	100.0
Pennsylvania	97.2	100.0	100.0	100.0	100.0
Rhode Island	82.8	98.7	100.0	100.0	100.0
South Carolina	94.6	96.5	100.0	100.0	97.7
Tennessee	94.7	99.6	100.0	100.0	99.3
Texas	94.5	100.0	100.0	100.0	100.0
Utah	90.2	98.7	99.1	98.8	98.5
Virginia	96.2	99.5	99.1	99.0	100.0
West Virginia	93.3	100.0	100.0	100.0	99.4
Wisconsin	89.8	98.0	100.0	100.0	97.7
Wyoming	88.8	100.0	94.5	98.1	100.0
TERRITORIES					
Guam	90.8	96.4	79.8	83.3	77.8
Virgin Islands	84.3	91.2	100.0	100.0	89.5

TABLE B.6 | Questionnaire Response Rates (continued)

PUBLIC SCHOOLS	Grade 8 - 1990				
	Weighted Percentage of Students Matched to Mathematics Teacher Questionnaires	Percentage of Mathematics Teacher Questionnaires Returned	Weighted Percentage of Students Matched to School Characteristics/ Policies Questionnaires	Percentage of School Characteristics/ Policies Questionnaires Returned	Percentage of Excluded Student Questionnaires Returned
NATION	76	72	88	84	90
Northeast	85	80	94	88	100
Southeast	78	73	91	87	85
Central	79	80	70	75	79
West	77	72	88	88	87
STATES					
Alabama	94	91	100	100	100
Arizona	85	84	99	99	98
Arkansas	92	90	98	98	100
California	86	86	98	97	95
Colorado	85	87	99	99	100
Connecticut	89	87	99	99	96
Delaware	85	83	98	97	98
Dist. Columbia	94	87	98	97	99
Florida	88	86	98	97	97
Georgia	87	89	98	98	100
Hawaii	91	88	99	98	99
Idaho	87	87	98	99	97
Indiana	87	86	96	98	93
Iowa	89	90	99	99	99
Kentucky	83	89	100	100	100
Louisiana	90	86	99	99	100
Maine	xxx	xxx	xxx	xxx	xxx
Maryland	89	90	99	99	96
Massachusetts	xxx	xxx	xxx	xxx	xxx
Michigan	91	91	100	100	99
Minnesota	88	86	99	99	97
Mississippi	xxx	xxx	xxx	xxx	xxx
Missouri	xxx	xxx	xxx	xxx	xxx
Nebraska	89	88	99	99	99
New Hampshire	88	83	100	100	99
New Jersey	91	88	99	98	97
New Mexico	90	88	97	97	93
New York	85	86	98	99	98
North Carolina	91	90	98	98	97
North Dakota	94	87	95	97	98
Ohio	83	83	100	100	98
Oklahoma	91	91	99	99	99
Pennsylvania	87	85	98	98	97
Rhode Island	87	84	100	100	100
South Carolina	xxx	xxx	xxx	xxx	xxx
Tennessee	xxx	xxx	xxx	xxx	xxx
Texas	84	89	99	99	99
Utah	xxx	xxx	xxx	xxx	xxx
Virginia	91	93	98	97	99
West Virginia	91	88	99	99	100
Wisconsin	87	81	99	96	98
Wyoming	84	81	100	98	99
TERRITORIES					
Guam	98	85	100	100	100
Virgin Islands	88	85	100	100	100

(xxx) Did not participate in the 1990 Trial State Assessment.

APPENDIX C

CONDITIONING VARIABLES AND CONTRAST CODINGS

APPENDIX C

Conditioning Variables and Contrast Codings

This appendix contains information about the conditioning variables used in the construction of plausible values for the 1992 Trial State Assessment Program in mathematics. Separate sets of conditioning variables were defined for each grade, but for the majority of the demographic and subject area background questions that were identical for grades four and eight, conditioning variables were similarly constructed for consistency. For both grades, two kinds of conditioning variables were defined—continuous or quasi-continuous variables, such as school mathematics score or number of hours spent watching television, and categorical variables which made up the majority of the conditioning variables created from responses to student, teacher, and school demographic and background questionnaires.

Categorical conditioning variables derived from questionnaire or demographic variables were incorporated into the conditioning process by constructing a set of contrasts, each of which defines one or more of the variable's response options. A recoding procedure explodes the raw student responses into a binary series of one-degree-of-freedom "dummy" variables. Questionnaire or demographic variables that possess ordinal response options, such as number of hours spent watching television, were included in the conditioning process by creating linear and/or quadratic multi-degree-of-freedom contrasts. Continuous variables were included in the conditioning process in their original form.

The remainder of this appendix gives the specifications used for constructing the conditioning variables. Table C-1 defines the information provided for each variable.

As described in Chapter 9, the linear conditioning model employed for the estimation of plausible values within each grade in each jurisdiction did not directly use the conditioning variable specifications listed in this appendix. To eliminate inherent instabilities in estimation encountered when using a large number of correlated variables, a principal component transformation of the correlation matrix obtained from the conditioning variable contrasts derived according to these primary specifications was performed. The principal components scores based on this transformation were used as the predictor variables in estimating the linear conditioning model.

Table C-1
Description of Data Provided for Each Conditioning Variable

Title	Description
CONDITIONING ID	An unique eight-character ID assigned to identify each conditioning variable corresponding to a particular background or subject area question within the entire pool of conditioning variables. The first four characters identify the origin of the variable: BACK (background questionnaire), MATH (student mathematics questionnaire), SCHL (school questionnaire), TCHR (teacher questionnaire), and TMAT (teacher mathematics questionnaire). The second four digits represent the sequential position within each origin group.
DESCRIPTION	A short description of the conditioning variable.
GRADES/ASSESSMENTS	Three characters identifying assessment ("S" for state, "N" for national) and grade (04, 08, and 12) in which the conditioning variable was used.
GROUP LABEL	A descriptive eight-character label identifying the conditioning variable.
NAEP ID	The seven-character NAEP database identification for the conditioning variable.
TYPE OF CONTRAST	The type of conditioning variable. "CLASS" identifies a categorical conditioning variable and "SCALE" identifies continuous or quasi-continuous conditioning variables.
LENGTH OF CONTRAST FIELD	The number of columns (or length of the contrast field) for the conditioning variable within the entire conditioning variable vector. The length is associated with the number of explicit contrasts comprising categorical conditioning variables.
DEGREES OF FREEDOM	The number of degrees of freedom for each contrast constructed for the conditioning variable.
NUMBER OF SPECIFICATION RECORDS	The number of unique contrasts corresponding to each conditioning variable. For each contrast a specifications record is given with the following information: a sequential identification number, an eight-character descriptive label corresponding to the associated questionnaire option(s), a "collapsing code string" enclosed in parentheses specifying the database values to be merged to form the contrast, the contrast itself, and a short description of the contrast.

```
CONDITIONING ID:          BACK0001
DESCRIPTION:              GRAND MEAN
GRADES/ASSESSMENTS:       W04, S04, W08, S08, W12
GROUP LABEL:              OVERALL          LENGTH OF CONTRAST FIELD       :  1
NAEP ID:                  BKSER            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         SCALE            NUMBER OF SPECIFICATION RECORDS:  1

001   OVERALL    (@          )  1          GRAND MEAN

CONDITIONING ID:          BACK0002
DESCRIPTION:              GENDER (DERIVED)
GRADES/ASSESSMENTS:       W04, S04, W08, S08, W12
GROUP LABEL:              GENDER           LENGTH OF CONTRAST FIELD       :  1
NAEP ID:                  DSEX             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS            NUMBER OF SPECIFICATION RECORDS:  2

001   MALE       (1          )  0          GENDER:  MALE
002   FEMALE     (2          )  1          GENDER:  FEMALE

CONDITIONING ID:          BACK0003
DESCRIPTION:              ETHNICITY/RACE (DERIVED)
GRADES/ASSESSMENTS:       W04, S04, W08, S08, W12
GROUP LABEL:              ETHNICTY         LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                  DRACE            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS            NUMBER OF SPECIFICATION RECORDS:  4

001   WHIT/AOM   (1,5,6,M    )  000        ETHNICITY:  WHITE, AMERICAN INDIAN, UNCLASSIFIED, MISSING
002   BLACK      (2          )  100        ETHNICITY:  BLACK
003   HISPANIC   (3          )  010        ETHNICITY:  HISPANIC
004   ASIAN      (4          )  001        ETHNICITY:  ASIAN AMERICAN

CONDITIONING ID:          BACK0005
DESCRIPTION:              TYPE OF COMMUNITY (STATE ONLY)
GRADES/ASSESSMENTS:       S04, S08
GROUP LABEL:              TOC              LENGTH OF CONTRAST FIELD       :  2
NAEP ID:                  TOC              DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS            NUMBER OF SPECIFICATION RECORDS:  3

001   TOC-OTHR   (1,4,M      )  00         TOC:  EXTREME RURAL, OTHER, MISSING
002   LO_METRO   (2          )  10         TOC:  LOW METROPOLITAN
003   HI_METRO   (3          )  01         TOC:  HIGH METROPOLITAN

CONDITIONING ID:          BACK0007
DESCRIPTION:              PARENTS' HIGHEST LEVEL OF EDUCATION
GRADES/ASSESSMENTS:       W04, S04, W08, S08, W12
GROUP LABEL:              PARED            LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                  PARED            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS            NUMBER OF SPECIFICATION RECORDS:  5

001   <HI_SCH    (1          )  0000       PARED:  LESS THAN HIGH SCHOOL
002   HS_GRAD    (2          )  1000       PARED:  HIGH SCHOOL GRADUATE
003   POST_HS    (3          )  0100       PARED:  POST HIGH SCHOOL
004   COL_GRAD   (4          )  0010       PARED:  COLLEGE GRADUATE
005   PARED-?    (M,IDK      )  0001       PARED:  MISSING, I DON'T KNOW

CONDITIONING ID:          BACK0008
DESCRIPTION:              ITEMS IN THE HOME (NEWSPAPER, > 25 BOOKS, ENCYCLOPEDIA, MAGAZINES)
GRADES/ASSESSMENTS:       W04, S04, W08, S08, W12
GROUP LABEL:              HOMEITMS         LENGTH OF CONTRAST FIELD       :  2
NAEP ID:                  HOMEEN2          DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS            NUMBER OF SPECIFICATION RECORDS:  3

001   HITEM<=2   (1,M        )  00         ITEMS IN HOME:  ZERO TO TWO ITEMS, MISSING
002   HITEM=3    (2          )  10         ITEMS IN HOME:  THREE ITEMS
003   HITEM=4    (3          )  01         ITEMS IN HOME:  FOUR ITEMS

CONDITIONING ID:          BACK0009
DESCRIPTION:              HOURS OF TV WATCHING (LINEAR)
GRADES/ASSESSMENTS:       W04, S04, W08, S08, W12
```

BEST COPY AVAILABLE

```
GROUP LABEL:              TVWATCHL          LENGTH OF CONTRAST FIELD       :  1
NAEP ID:                  B001801           DEGREES OF FREEDOM PER CONTRAST:  6
TYPE OF CONTRAST:         SCALE             NUMBER OF SPECIFICATION RECORDS:  7

001   TV-LIN1    (1          )  0           TV WATCHING (LINEAR):  NONE
002   TV-LIN2    (2          )  1           TV WATCHING (LINEAR):  ONE HOUR OR LESS PER DAY
003   TV-LIN3    (3          )  2           TV WATCHING (LINEAR):  TWO HOURS PER DAY
004   TV-LIN4    (4,M        )  3           TV WATCHING (LINEAR):  THREE HOURS PER DAY
005   TV-LIN5    (5          )  4           TV WATCHING (LINEAR):  FOUR HOURS PER DAY
006   TV-LIN6    (6          )  5           TV WATCHING (LINEAR):  FIVE HOURS PER DAY
007   TV-LIN7    (7          )  6           TV WATCHING (LINEAR):  SIX OR MORE HOURS PER DAY

CONDITIONING ID:          BACK0010
DESCRIPTION:              HOURS OF TV WATCHING (QUADRATIC)
GRADES/ASSESSMENTS:       N04, S04, N08, S08, N12
GROUP LABEL:              TVWATCHQ          LENGTH OF CONTRAST FIELD       :  2
NAEP ID:                  B001801           DEGREES OF FREEDOM PER CONTRAST:  6
TYPE OF CONTRAST:         SCALE             NUMBER OF SPECIFICATION RECORDS:  7

001   TV-QUAD1   (1          )  00          TV WATCHING (QUADRATIC):  NONE
002   TV-QUAD2   (2          )  01          TV WATCHING (QUADRATIC):  ONE HOUR OR LESS PER DAY
003   TV-QUAD3   (3          )  04          TV WATCHING (QUADRATIC):  TWO HOURS PER DAY
004   TV-QUAD4   (4,M        )  09          TV WATCHING (QUADRATIC):  THREE HOURS PER DAY
005   TV-QUAD5   (5          )  16          TV WATCHING (QUADRATIC):  FOUR HOURS PER DAY
006   TV-QUAD6   (6          )  25          TV WATCHING (QUADRATIC):  FIVE HOURS PER DAY
007   TV-QUAD7   (7          )  36          TV WATCHING (QUADRATIC):  SIX OR MORE HOURS PER DAY

CONDITIONING ID:          BACK0011
DESCRIPTION:              HOME LANG MINORITY (HOW OFTEN DO PEOPLE IN HOME SPEAK LANG OTHER THAN ENGLISH?)
GRADES/ASSESSMENTS:       N04, S04, N08, S08, N12
GROUP LABEL:              HOMELANG          LENGTH OF CONTRAST FIELD       :  1
NAEP ID:                  B003201           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS             NUMBER OF SPECIFICATION RECORDS:  2

001   HL-NEV/?   (1,M        )  0           HOME LANGUAGE MINORITY:  NEVER, MISSING
002   HL-SM/AL   (2,3        )  1           HOME LANGUAGE MINORITY:  SOMTIMES, ALWAYS

CONDITIONING ID:          BACK0012
DESCRIPTION:              HOMEWORK ASSIGNED? (GRADE 4)
GRADES/ASSESSMENTS:       N04, S04
GROUP LABEL:              HW-CORE4          LENGTH OF CONTRAST FIELD       :  2
NAEP ID:                  B006601           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS             NUMBER OF SPECIFICATION RECORDS:  3

001   HW4-MISS   (M          )  00          HOMEWORK ASSIGNED?:  MISSING
002   HW4-NONE   (1          )  10          HOMEWORK ASSIGNED?:  NO HOMEWORK ASSIGNED
003   HW4-YES    (2-5        )  01          HOMEWORK ASSIGNED?:  YES

CONDITIONING ID:          BACK0013
DESCRIPTION:              AMOUNT OF HOMEWORK (LINEAR) (GRADE 4)
GRADES/ASSESSMENTS:       N04, S04
GROUP LABEL:              HWORKL4           LENGTH OF CONTRAST FIELD       :  1
NAEP ID:                  B006601           DEGREES OF FREEDOM PER CONTRAST:  3
TYPE OF CONTRAST:         SCALE             NUMBER OF SPECIFICATION RECORDS:  4

001   HW4-LIN1   (1,2,M      )  0           AMOUNT OF HOMEWORK (LINEAR):  DON'T HAVE, DON'T DO, MISSING
002   HW4-LIN2   (3          )  1           AMOUNT OF HOMEWORK (LINEAR):  ONE HALF HOUR
003   HW4-LIN3   (4          )  2           AMOUNT OF HOMEWORK (LINEAR):  ONE HOUR
004   HW4-LIN4   (5          )  3           AMOUNT OF HOMEWORK (LINEAR):  MORE THAN ONE HOUR


CONDITIONING ID:          BACK0014
DESCRIPTION:              AMOUNT OF HOMEWORK (QUADRATIC) (GRADE 4)
GRADES/ASSESSMENTS:       N04, S04
GROUP LABEL:              HWORKQ4           LENGTH OF CONTRAST FIELD       :  1
NAEP ID:                  B006601           DEGREES OF FREEDOM PER CONTRAST:  3
TYPE OF CONTRAST:         SCALE             NUMBER OF SPECIFICATION RECORDS:  4

001   HW4QUAD1   (1,2,M      )  0           AMOUNT OF HOMEWORK (QUADRATIC):  DON'T HAVE ANY, DON'T DO
```

```
ANY, MISSING
002   HW4QUAD2   (3           )  1          AMOUNT OF HOMEWORK (QUADRATIC):  ONE HALF HOUR
003   HW4QUAD3   (4           )  4          AMOUNT OF HOMEWORK (QUADRATIC):  ONE HOUR
004   HW4QUAD4   (5           )  9          AMOUNT OF HOMEWORK (QUADRATIC):  MORE THAN ONE HOUR

CONDITIONING ID:          BACK0015
DESCRIPTION:              HOMEWORK ASSIGNED?
GRADES/ASSESSMENTS:       N08, S08, N12
GROUP LABEL:              HW-CORE            LENGTH OF CONTRAST FIELD       :   2
NAEP ID:                  B003901            DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:         CLASS              NUMBER OF SPECIFICATION RECORDS:   3

001   HWC-MISS   (M           )  00         HOMEWORK ASSIGNED?:  MISSING
002   HWC-NONE   (1           )  10         HOMEWORK ASSIGNED?:  NO HOMEWORK ASSIGNED
003   HWC-YES    (2-6         )  01         HOMEWORK ASSIGNED?:  YES

CONDITIONING ID:          BACK0016
DESCRIPTION:              AMOUNT OF HOMEWORK (LINEAR)
GRADES/ASSESSMENTS:       N08, S08, N12
GROUP LABEL:              HMWRKL             LENGTH OF CONTRAST FIELD       :   1
NAEP ID:                  B003901            DEGREES OF FREEDOM PER CONTRAST:   4
TYPE OF CONTRAST:         SCALE              NUMBER OF SPECIFICATION RECORDS:   5

001   HW-LIN1    (1,2,M       )  0          AMOUNT OF HOMEWORK (LINEAR):  DON'T HAVE, DON'T DO, MISSING
002   HW-LIN2    (3           )  1          AMOUNT OF HOMEWORK (LINEAR):  ONE HALF HOUR
003   HW-LIN3    (4           )  2          AMOUNT OF HOMEWORK (LINEAR):  ONE HOUR
004   HW-LIN4    (5           )  3          AMOUNT OF HOMEWORK (LINEAR):  TWO HOURS
005   HW-LIN5    (6           )  4          AMOUNT OF HOMEWORK (LINEAR):  MORE THAN TWO HOURS

CONDITIONING ID:          BACK0017
DESCRIPTION:              AMOUNT OF HOMEWORK (QUADRATIC)
GRADES/ASSESSMENTS:       N08, S08, N12
GROUP LABEL:              HMWRKQ             LENGTH OF CONTRAST FIELD       :   2
NAEP ID:                  B003901            DEGREES OF FREEDOM PER CONTRAST:   4
TYPE OF CONTRAST:         SCALE              NUMBER OF SPECIFICATION RECORDS:   5

001   HW-QUAD1   (1,2,M       )  00         AMOUNT OF HOMEWORK (QUADRATIC):  DON'T HAVE ANY, DON'T DO
ANY, MISSING
002   HW-QUAD2   (3           )  01         AMOUNT OF HOMEWORK (QUADRATIC):  ONE HALF HOUR
003   HW-QUAD3   (4           )  04         AMOUNT OF HOMEWORK (QUADRATIC):  ONE HOUR
004   HW-QUAD4   (5           )  09         AMOUNT OF HOMEWORK (QUADRATIC):  TWO HOURS
005   HW-QUAD5   (6           )  16         AMOUNT OF HOMEWORK (QUADRATIC):  MORE THAN TWO HOURS

CONDITIONING ID:          BACK0018
DESCRIPTION:              PERCENT WHITE STUDENTS IN SCHOOL
GRADES/ASSESSMENTS:       N04, S04, N08, S08, N12
GROUP LABEL:              %WHITE             LENGTH OF CONTRAST FIELD       :   2
NAEP ID:                  PCTWHT             DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:         CLASS              NUMBER OF SPECIFICATION RECORDS:   3

001   PREDOM/?   (80-110,M    )  00         PREDOMINANTLY WHITE, MISSING
002   MINORITY   (0-49        )  10         WHITE MINORITY
003   INTEGRAT   (50-79       )  01         INTEGRATED

CONDITIONING ID:          BACK0020
DESCRIPTION:              SCHOOL TYPE:  PUBLIC/NON-PUBLIC
GRADES/ASSESSMENTS:       N04, N08, N12
GROUP LABEL:              SCH_TYPE           LENGTH OF CONTRAST FIELD       :   1
NAEP ID:                  SCHTYPE            DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:         CLASS              NUMBER OF SPECIFICATION RECORDS:   2

001   PUBLIC     (1           )  0          PUBLIC SCHOOL
002   NON_PUBL   (2-5,M       )  1          PRIVATE, CATHOLIC, BUREAU OF INDIAN AFFAIRS, DEPT OF DEFENSE,
MISSING

CONDITIONING ID:          BACK0021
DESCRIPTION:              SINGLE/MULTIPLE PARENT(S) AT HOME
GRADES/ASSESSMENTS:       N04, S04, N08, S08, N12
GROUP LABEL:              PARENTS            LENGTH OF CONTRAST FIELD       :   1
```

```
NAEP ID:                  SINGLEP           DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:         CLASS             NUMBER OF SPECIFICATION RECORDS:   2

001   NOT2PARS   (2-4,M       )  0          NOT TWO PARENTS, MISSING
002   2PARENTS   (1           )  1          BOTH FATHER AND MOTHER AT HOME

CONDITIONING ID:          BACK0022
DESCRIPTION:              MOTHER AT HOME
GRADES/ASSESSMENTS:       N04, S04, N08, S08, N12
GROUP LABEL:              MOM@HOME          LENGTH OF CONTRAST FIELD       :   1
NAEP ID:                  B005601           DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:         CLASS             NUMBER OF SPECIFICATION RECORDS:   2

001   MOM@HM-N   (2,M         )  0          MOTHER AT HOME:  NO, MISSING
002   MOM@HM-Y   (1           )  1          MOTHER AT HOME:  YES

CONDITIONING ID:          BACK0023
DESCRIPTION:              PAGES READ FOR SCHOOL AND HOMEWORK EACH DAY (CONTRAST 1)
GRADES/ASSESSMENTS:       N04, S04, N08, S08, N12
GROUP LABEL:              PGSREAD1          LENGTH OF CONTRAST FIELD       :   1
NAEP ID:                  B001101           DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:         CLASS             NUMBER OF SPECIFICATION RECORDS:   2

001   <=5_PGS    (5,M         )  0          PAGES READ (1):  15 OR FEWER PAGES, MISSING
002   >=6_PGS    (1-4         )  1          PAGES READ (1):  > 20 PGS, 16-20 PGS, 11-15 PGS, 6-10 PGS

CONDITIONING ID:          BACK0024
DESCRIPTION:              PAGES READ FOR SCHOOL AND HOMEWORK EACH DAY (CONTRAST 2)
GRADES/ASSESSMENTS:       N04, S04, N08, S08, N12
GROUP LABEL:              PGSREAD2          LENGTH OF CONTRAST FIELD       :   1
NAEP ID:                  B001101           DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:         CLASS             NUMBER OF SPECIFICATION RECORDS:   2

001   <=10_PGS   (4,5,M       )  0          PAGES READ (2):  6-10 PAGES, 5 OR FEWER PAGES, MISSING
002   >=11_PGS   (1-3         )  1          PAGES READ (2):  MORE THAN 20 PAGES, 16-20 PAGES, 11-15 PAGES

CONDITIONING ID:          BACK0025
DESCRIPTION:              WENT TO PRESCHOOL?
GRADES/ASSESSMENTS:       N04, S04
GROUP LABEL:              PRESCH            LENGTH OF CONTRAST FIELD       :   1
NAEP ID:                  B004201           DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:         CLASS             NUMBER OF SPECIFICATION RECORDS:   2

001   PRESCH-N   (2,3,IDK,M   )  0          WENT TO PRESCHOOL?:  NO, I DON'T KNOW, MISSING
002   PRESCH-Y   (1           )  1          WENT TO PRESCHOOL?:  YES

CONDITIONING ID:          BACK0026
DESCRIPTION:              DAYS OF SCHOOL MISSED LAST MONTH
GRADES/ASSESSMENTS:       N08, S08, N12
GROUP LABEL:              SCH_MISS          LENGTH OF CONTRAST FIELD       :   1
NAEP ID:                  S004001           DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:         CLASS             NUMBER OF SPECIFICATION RECORDS:   2

001   >=3_DAYS   (3-5,M       )  0          DAYS SCHOOL MISSED:  3 OR 4 DAYS, 5-10 DAYS, > 10 DAYS, MSSNG
002   <=2_DAYS   (1,2         )  1          DAYS SCHOOL MISSED:  NONE, 1 OR 2 DAYS

CONDITIONING ID:          BACK0042
DESCRIPTION:              BORN IN ONE OF THE 50 STATES
GRADES/ASSESSMENTS:       N04, S04, N08, S08, N12
GROUP LABEL:              BORN_USA          LENGTH OF CONTRAST FIELD       :   1
NAEP ID:                  B007801           DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:         CLASS             NUMBER OF SPECIFICATION RECORDS:   2

001   USA-YES    (1           )  0          BORN IN THE USA:  YES
002   USA-NO/?   (2,M         )  1          BORN IN THE USA:  NO/MIS SING

CONDITIONING ID:          BACK0043
DESCRIPTION:              HOW MANY TIMES CHANGED SCHOOLS IN THE LAST TWO YEARS?
GRADES/ASSESSMENTS:       N04, S04, N08, S08
```

```
GROUP LABEL:              SCH_CHGS           LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                  B007301            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS              NUMBER OF SPECIFICATION RECORDS:  4

001   CHGSCH=0   (1          )  000          CHANGED SCHOOLS (NONE)
002   CHGSCH=1   (2          )  100          CHANGED SCHOOLS ONCE
003   CHGSCH=2   (3          )  010          CHANGED SCHOOLS TWICE
004   CHGSCH3+   (4,M        )  001          CHANGED SCHOOLS 3 OR MORE TIMES, MISSING

CONDITIONING ID:          BACK0044
DESCRIPTION:              HOW MANY GRADES HAVE YOU GONE TO SCHOOL IN THIS STATE? (K-4)
GRADES/ASSESSMENTS:       N04, S04
GROUP LABEL:              GRDS_ST4           LENGTH OF CONTRAST FIELD       :  2
NAEP ID:                  B007601            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS              NUMBER OF SPECIFICATION RECORDS:  3

001   ST4GRD<1   (1,M        )  00           LESS THAN ONE GRADE IN THIS STATE, MISSING (K-4)
002   ST4GRD12   (2          )  10           ONE TO TWO GRADES IN THIS STATE (K-4)
003   ST4GRD3+   (3          )  01           THREE OR MORE GRADES IN THIS STATE (K-4)

CONDITIONING ID:          BACK0045
DESCRIPTION:              HOW OFTEN DO YOU DISCUSS THINGS STUDIED IN SCHOOL WITH SOMEONE AT HOME?
GRADES/ASSESSMENTS:       N04, S04, N08, S08, N12
GROUP LABEL:              DISC@HOM           LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                  B007401            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS              NUMBER OF SPECIFICATION RECORDS:  4

001   DIS@HOM1   (1          )  000          DISCUSS AT HOME (ALMOST EVERYDAY)
002   DIS@HOM2   (2          )  100          DISCUSS AT HOME (ONCE OR TWICE A WEEK)
003   DIS@HOM3   (3          )  010          DISCUSS AT HOME (ONCE OR TWICE A MONTH)
004   DIS@HOM4   (4,M        )  001          DISCUSS AT HOME (NEVER OR HARDLY EVER, MISSING)

CONDITIONING ID:          BACK0046
DESCRIPTION:              HOW OFTEN DO USE A COMPUTER FOR SCHOOLWORK?
GRADES/ASSESSMENTS:       N04, S04, N08, S08, N12
GROUP LABEL:              COMP4SCH           LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                  B007501            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS              NUMBER OF SPECIFICATION RECORDS:  5

001   CMP4SCH1   (1          )  0000         COMPUTER FOR SCHOOLWORK (ALMOST EVERYDAY)
002   CMP4SCH2   (2          )  1000         COMPUTER FOR SCHOOLWORK (ONCE OR TWICE A WEEK)
003   CMP4SCH3   (3          )  0100         COMPUTER FOR SCHOOLWORK (ONCE OR TWICE A MONTH)
004   CMP4SCH4   (4          )  0010         COMPUTER FOR SCHOOLWORK (NEVER OR HARDLY EVER)
005   CMP4SCH5   (M          )  0001         COMPUTER FOR SCHOOLWORK (MISSING)


CONDITIONING ID:          BACK0047
DESCRIPTION:              HOW MANY GRADES HAVE YOU GONE TO SCHOOL IN THIS STATE? (K-8)
GRADES/ASSESSMENTS:       N08, S08
GROUP LABEL:              GRDS_ST8           LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                  B007701            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS              NUMBER OF SPECIFICATION RECORDS:  4

001   ST8GRD<1   (1,M        )  000          LESS THAN ONE GRADE IN THIS STATE, MISSING (K-8)
002   ST8GRD12   (2          )  100          ONE TO TWO GRADES IN THIS STATE (K-8)
003   ST8GRD35   (3          )  010          THREE TO FIVE GRADES IN THIS STATE (K-8)
004   ST8GRD>5   (4          )  001          MORE THAN FIVE GRADES IN THIS STATE (K-8)

CONDITIONING ID:          MATH0001
DESCRIPTION:              SCHOOL LEVEL AVERAGE MATHEMATICS PROFICIENCY
GRADES/ASSESSMENTS:       N04, S04, N08, S08, N12
GROUP LABEL:              SLP_MATH           LENGTH OF CONTRAST FIELD       :  1
NAEP ID:                  SCHMATH            DEGREES OF FREEDOM PER CONTRAST: 999
TYPE OF CONTRAST:         CLASS              NUMBER OF SPECIFICATION RECORDS:  2

001   SLP_MA-Y   (@          )  1            SCHOOL LEVEL AVERAGE MATHEMATICS PROFICIENCY NOT-MISSING
002   SLP_MA-?   (M          )  0            SCHOOL LEVEL AVERAGE MATHEMATICS PROFICIENCY MISSING

CONDITIONING ID:          MATH0002
```

```
DESCRIPTION:              SCHOOL LEVEL AVERAGE MATHEMATICS PROFICIENCY
GRADES/ASSESSMENTS:       N04, S04, N08, S08, N12
GROUP LABEL:              SLP_MATL            LENGTH OF CONTRAST FIELD       :  8
NAEP ID:                  SCHMATH             DEGREES OF FREEDOM PER CONTRAST: 999
TYPE OF CONTRAST:         SCALE               NUMBER OF SPECIFICATION RECORDS:  2

001   SLP_MA-L   (#           )  (F8.4)      SCHOOL LEVEL AVERAGE MATHEMATICS PROFICIENCY MEAN
002   SLP_MA-?   (M           )  0           SCHOOL LEVEL AVERAGE MATHEMATICS PROFICIENCY MISSING

CONDITIONING ID:          MATH0003
DESCRIPTION:              HOW OFTEN DO MATH PROBLEMS FROM TEXTBOOKS (STUDENT)?
GRADES/ASSESSMENTS:       N04, S04, N08, S08, N12
GROUP LABEL:              S_TXTBKS            LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                  M811601             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS               NUMBER OF SPECIFICATION RECORDS:  5

001   S_TXTBK1   (1           )  0000        MATH FROM TEXTBOOKS (STUDENT):  ALMOST EVERY DAY
002   S_TXTBK2   (2           )  1000        MATH FROM TEXTBOOKS (STUDENT):  ONCE OR TWICE A WEEK
003   S_TXTBK3   (3           )  0100        MATH FROM TEXTBOOKS (STUDENT):  ONCE OR TWICE A MONTH
004   S_TXTBK4   (4           )  0010        MATH FROM TEXTBOOKS (STUDENT):  NEVER OR HARDLY EVER
005   S_TXTBK?   (M           )  0001        MATH FROM TEXTBOOKS (STUDENT):  MISSING

CONDITIONING ID:          MATH0004
DESCRIPTION:              HOW OFTEN DO MATH PROBLEMS ON WORKSHEETS?
GRADES/ASSESSMENTS:       N04, S04, N08, S08, N12
GROUP LABEL:              S_WRKSHS            LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                  M811602             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS               NUMBER OF SPECIFICATION RECORDS:  5

001   S_WRKSH1   (1           )  0000        MATH FROM WORKSHEETS (STUDENT): ALMOST EVERY DAY
002   S_WRKSH2   (2           )  1000        MATH FROM WORKSHEETS (STUDENT): ONCE OR TWICE A WEEK
003   S_WRKSH3   (3           )  0100        MATH FROM WORKSHEETS (STUDENT): ONCE OR TWICE A MONTH
004   S_WRKSH4   (4           )  0010        MATH FROM WORKSHEETS (STUDENT): NEVER OR HARDLY EVER
005   S_WRKSH?   (M           )  0001        MATH FROM WORKSHEETS (STUDENT): MISSING


CONDITIONING ID:          MATH0005
DESCRIPTION:              HOW OFTEN SOLVE MATH PROBLEMS IN SMALL GROUPS?
GRADES/ASSESSMENTS:       N04, S04, N08, S08, N12
GROUP LABEL:              S_SMGRPS            LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                  M811603             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS               NUMBER OF SPECIFICATION RECORDS:  5

001   S_SMGRP1   (1           )  0000        MATH IN SMALL GROUPS (STUDENT): ALMOST EVERY DAY
002   S_SMGRP2   (2           )  1000        MATH IN SMALL GROUPS (STUDENT): ONCE OR TWICE A WEEK
003   S_SMGRP3   (3           )  0100        MATH IN SMALL GROUPS (STUDENT): ONCE OR TWICE A MONTH
004   S_SMGRP4   (4           )  0010        MATH IN SMALL GROUPS (STUDENT): NEVER OR HARDLY EVER
005   S_SMGRP?   (M           )  0001        MATH IN SMALL GROUPS (STUDENT): MISSING

CONDITIONING ID:          MATH0006
DESCRIPTION:              HOW OFTEN WORK WITH OBJECTS LIKE RULERS, BLOCKS, SHAPES? (STUDENT)
GRADES/ASSESSMENTS:       N04, S04
GROUP LABEL:              S_OBJCTS            LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                  M811604             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS               NUMBER OF SPECIFICATION RECORDS:  5

001   S_OBJCT1   (1           )  0000        WORK WITH OBJECTS (STUDENT): ALMOST EVERY DAY
002   S_OBJCT2   (2           )  1000        WORK WITH OBJECTS (STUDENT): ONCE OR TWICE A WEEK
003   S_OBJCT3   (3           )  0100        WORK WITH OBJECTS (STUDENT): ONCE OR TWICE A MONTH
004   S_OBJCT4   (4           )  0010        WORK WITH OBJECTS (STUDENT): NEVER OR HARDLY EVER
005   S_OBJCT?   (M           )  0001        WORK WITH OBJECTS (STUDENT): MISSING

CONDITIONING ID:          MATH0007
DESCRIPTION:              HOW OFTEN WORK WITH MEASUREMENT INSTRUMENTS/GEOMETRIC SOLIDS? (STUDENT)
GRADES/ASSESSMENTS:       N08, S08, N12
GROUP LABEL:              S_MI&GS             LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                  M811608             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS               NUMBER OF SPECIFICATION RECORDS:  5
```

```
001   S_MI&GS1   (1           )  0000        MEASUREMENT INSTR/GEOM SOLIDS (STUDENT): ALMOST EVERY DAY
002   S_MI&GS2   (2           )  1000        MEASUREMENT INSTR/GEOM SOLIDS (STUDENT): ONCE OR TWICE A WEEK
003   S_MI&GS3   (3           )  0100        MEASUREMENT INSTR/GEOM SOLIDS (STUDENT): ONCE OR TWICE MONTH
004   S_MI&GS4   (4           )  0010        MEASUREMENT INSTR/GEOM SOLIDS (STUDENT): NEVER OR HARDLY EVER
005   S_MI&GS?   (M           )  0001        MEASUREMENT INSTR/GEOM SOLIDS (STUDENT): MISSING

CONDITIONING ID:           MATH0008
DESCRIPTION:               HOW OFTEN USE A CALCULATOR (STUDENT)?
GRADES/ASSESSMENTS:        N04, S04, N08, S08, N12
GROUP LABEL:               S_CALCTR          LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                   M811605           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:          CLASS             NUMBER OF SPECIFICATION RECORDS:  5

001   S_CALCR1   (1           )  0000        USE CALCULATOR (STUDENT): ALMOST EVERY DAY
002   S_CALCR2   (2           )  1000        USE CALCULATOR (STUDENT): ONCE OR TWICE A WEEK
003   S_CALCR3   (3           )  0100        USE CALCULATOR (STUDENT): ONCE OR TWICE A MONTH
004   S_CALCR4   (4           )  0010        USE CALCULATOR (STUDENT): NEVER OR HARDLY EVER
005   S_CALCR?   (M           )  0001        USE CALCULATOR (STUDENT): MISSING

CONDITIONING ID:           MATH0009
DESCRIPTION:               HOW OFTEN USE A COMPUTER (STUDENT)?
GRADES/ASSESSMENTS:        N04, S04, N08, S08, N12
GROUP LABEL:               S_COMPTR          LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                   M811606           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:          CLASS             NUMBER OF SPECIFICATION RECORDS:  5

001   S_CMPTR1   (1           )  0000        USE COMPUTER (STUDENT): ALMOST EVERY DAY
002   S_CMPTR2   (2           )  1000        USE COMPUTER (STUDENT): ONCE OR TWICE A WEEK
003   S_CMPTR3   (3           )  0100        USE COMPUTER (STUDENT): ONCE OR TWICE A MONTH
004   S_CMPTR4   (4           )  0010        USE COMPUTER (STUDENT): NEVER OR HARDLY EVER
005   S_CMPTR?   (M           )  0001        USE COMPUTER (STUDENT): MISSING

CONDITIONING ID:           MATH0010
DESCRIPTION:               HOW OFTEN WRITE ABOUT SOLVING MATH PROBLEM (STUDENT)?
GRADES/ASSESSMENTS:        N08, S08, N12
GROUP LABEL:               S_PRBSOL          LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                   M811609           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:          CLASS             NUMBER OF SPECIFICATION RECORDS:  5

001   S_PRBSL1   (1           )  0000        PROBLEM SOLVING (STUDENT): ALMOST EVERY DAY
002   S_PRBSL2   (2           )  1000        PROBLEM SOLVING (STUDENT): ONCE OR TWICE A WEEK
003   S_PRBSL3   (3           )  0100        PROBLEM SOLVING (STUDENT): ONCE OR TWICE A MONTH
004   S_PRBSL4   (4           )  0010        PROBLEM SOLVING (STUDENT): NEVER OR HARDLY EVER
005   S_PRBSL?   (M           )  0001        PROBLEM SOLVING (STUDENT): MISSING

CONDITIONING ID:           MATH0011
DESCRIPTION:               HOW OFTEN MAKE UP MATH PROBLEMS FOR OTHERS TO SOLVE? (STUDENT)
GRADES/ASSESSMENTS:        N08, S08, N12
GROUP LABEL:               S_MUPROB          LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                   M811610           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:          CLASS             NUMBER OF SPECIFICATION RECORDS:  5

001   S_MUPRB1   (1           )  0000        STUDENTS MAKE UP PROBLEMS (STUDENT): ALMOST EVERY DAY
002   S_MUPRB2   (2           )  1000        STUDENTS MAKE UP PROBLEMS (STUDENT): ONCE OR TWICE A WEEK
003   S_MUPRB3   (3           )  0100        STUDENTS MAKE UP PROBLEMS (STUDENT): ONCE OR TWICE A MONTH
004   S_MUPRB4   (4           )  0010        STUDENTS MAKE UP PROBLEMS (STUDENT): NEVER OR HARDLY EVER
005   S_MUPRB?   (M           )  0001        STUDENTS MAKE UP PROBLEMS (STUDENT): MISSING

CONDITIONING ID:           MATH0012
DESCRIPTION:               HOW OFTEN TAKE MATH TESTS? (STUDENT)
GRADES/ASSESSMENTS:        N04, S04, N08, S08, N12
GROUP LABEL:               S_MATTST          LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                   M811607           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:          CLASS             NUMBER OF SPECIFICATION RECORDS:  5

001   S_MATST1   (1           )  0000        TAKE MATH TESTS (STUDENT): ALMOST EVERY DAY
002   S_MATST2   (2           )  1000        TAKE MATH TESTS (STUDENT): ONCE OR TWICE A WEEK
003   S_MATST3   (3           )  0100        TAKE MATH TESTS (STUDENT): ONCE OR TWICE A MONTH
004   S_MATST4   (4           )  0010        TAKE MATH TESTS (STUDENT): NEVER OR HARDLY EVER
```

```
005   S_MATST?   (M            )  0001         TAKE MATH TESTS (STUDENT): MISSING

CONDITIONING ID:            MATH0013
DESCRIPTION:                HOW OFTEN WRITE REPORTS/DO PROJECTS? (STUDENT)
GRADES/ASSESSMENTS:         N08, S08, N12
GROUP LABEL:                S_REPPRJ          LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                    M811611           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS             NUMBER OF SPECIFICATION RECORDS:  5

001   S_REPPJ1   (1            )  0000         REPORTS/PROJECTS (STUDENT): ALMOST EVERY DAY
002   S_REPPJ2   (2            )  1000         REPORTS/PROJECTS (STUDENT): ONCE OR TWICE A WEEK
003   S_REPPJ3   (3            )  0100         REPORTS/PROJECTS (STUDENT): ONCE OR TWICE A MONTH
004   S_REPPJ4   (4            )  0010         REPORTS/PROJECTS (STUDENT): NEVER OR HARDLY EVER
005   S_REPPJ?   (M            )  0001         REPORTS/PROJECTS (STUDENT): MISSING

CONDITIONING ID:            MATH0014
DESCRIPTION:                HAVE A CALCULATOR TO DO MATH SCHOOLWORK?
GRADES/ASSESSMENTS:         N04, S04, N08, S08, N12
GROUP LABEL:                HAVECALC          LENGTH OF CONTRAST FIELD       :  1
NAEP ID:                    M811201           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS             NUMBER OF SPECIFICATION RECORDS:  2

001   HVCALC-Y   (1            )  0            HAVE CALCULATOR (STUDENT):  YES
002   HVCALC-N   (2,M          )  1            HAVE CALCULATOR (STUDENT):  NO, MISSING

CONDITIONING ID:            MATH0015
DESCRIPTION:                HAVE YOU EVER USED A SCIENTIFIC CALCULATOR?
GRADES/ASSESSMENTS:         N08, S08, N12
GROUP LABEL:                SCI_CALC          LENGTH OF CONTRAST FIELD       :  1
NAEP ID:                    M810401           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS             NUMBER OF SPECIFICATION RECORDS:  2

001   SCICAL-Y   (1            )  0            USED SCIENTIFIC CALCULATOR:  YES
002   SCICAL-N   (2,M          )  1            USED SCIENTIFIC CALCULATOR:  NO, MISSING

CONDITIONING ID:            MATH0016
DESCRIPTION:                WHAT KIND OF MATH CLASS ARE YOU TAKING THIS YEAR? (GRADE 8)
GRADES/ASSESSMENTS:         N08, S08
GROUP LABEL:                MCLASS8           LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                    M810501           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS             NUMBER OF SPECIFICATION RECORDS:  5

001   NO_MATH8   (1,M          )  0000         8TH GRADE MATH CLASS: NO MATH THIS YEAR, MISSING
002   8GR_MAT8   (2            )  1000         8TH GRADE MATH CLASS: 8TH GRADE MATH
003   PRE-ALG8   (3            )  0100         8TH GRADE MATH CLASS: PRE-ALGEBRA
004   ALGEBRA8   (4            )  0010         8TH GRADE MATH CLASS: ALGEBRA
005   OTHER_M8   (5            )  0001         8TH GRADE MATH CLASS: OTHER

CONDITIONING ID:            MATH0017
DESCRIPTION:                WHAT MATH CLASS WILL YOU TAKE IN 9TH GRADE? (GRADE 8)
GRADES/ASSESSMENTS:         N08, S08
GROUP LABEL:                MCLASS9           LENGTH OF CONTRAST FIELD       :  6
NAEP ID:                    M811701           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS             NUMBER OF SPECIFICATION RECORDS:  7

001   NO_MATH9   (1            )  000000       9TH GRADE MATH CLASS: WON'T TAKE MATH IN 9TH GRADE
002   BASIC9     (2            )  100000       9TH GRADE MATH CLASS: BASIC, GENERAL, BUSINESS, CONSUMER MATH
003   PRE-ALG9   (3            )  010000       9TH GRADE MATH CLASS: PRE-ALGEBRA
004   ALGEBRA9   (4            )  001000       9TH GRADE MATH CLASS: ALGEBRA
005   GEOMTRY9   (5            )  000100       9TH GRADE MATH CLASS: GEOMETRY
006   OTHER_M9   (6            )  000010       9TH GRADE MATH CLASS: OTHER
007   MCLASS9?   (7,M          )  000001       9TH GRADE MATH CLASS: MISSING

CONDITIONING ID:            MATH0020
DESCRIPTION:                HOW MUCH TIME SPENT ON MATH HOMEWORK EACH DAY? (STUDENT)
GRADES/ASSESSMENTS:         N04, S04, N08, S08, N12
GROUP LABEL:                S_MATHHW          LENGTH OF CONTRAST FIELD       :  7
NAEP ID:                    M811301           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS             NUMBER OF SPECIFICATION RECORDS:  8
```

```
001   S_MATHW1   (1          )  0000000      AMOUNT MATH HOMEWORK (STUDENT):  NOT TAKING MATH
002   S_MATHW1   (2          )  1000000      AMOUNT MATH HOMEWORK (STUDENT):  NONE
003   S_MATHW2   (3          )  0100000      AMOUNT MATH HOMEWORK (STUDENT):  15 MINUTES
004   S_MATHW3   (4          )  0010000      AMOUNT MATH HOMEWORK (STUDENT):  30 MINUTES
005   S_MATHW4   (5          )  0001000      AMOUNT MATH HOMEWORK (STUDENT):  45 MINUTES
006   S_MATHW5   (6          )  0000100      AMOUNT MATH HOMEWORK (STUDENT):  1 HOUR
007   S_MATHW6   (7          )  0000010      AMOUNT MATH HOMEWORK (STUDENT):  MORE THAN 1 HOUR
008   S_MATHW7   (M          )  0000001      AMOUNT MATH HOMEWORK (STUDENT):  MISSING, DOES NOT APPLY

CONDITIONING ID:              MATH0021
DESCRIPTION:                  DO YOU GET HELP IN MATH FROM SPECIAL TEACHERS?
GRADES/ASSESSMENTS:           N04, S04
GROUP LABEL:                  MATHHELP            LENGTH OF CONTRAST FIELD      :   1
NAEP ID:                      M811401             DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:             CLASS               NUMBER OF SPECIFICATION RECORDS:   2

001   MATHLP-Y   (1          )  0                HELP IN MATH:  YES
002   MATHLP-N   (2,M        )  1                HELP IN MATH:  NO, MISSING

CONDITIONING ID:              MATH0022
DESCRIPTION:                  AGREE/DISAGREE: I LIKE MATH (GRADE 4)
GRADES/ASSESSMENTS:           N04, S04
GROUP LABEL:                  LIKEMAT4            LENGTH OF CONTRAST FIELD      :   2
NAEP ID:                      M811101             DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:             CLASS               NUMBER OF SPECIFICATION RECORDS:   3

001   LKMATH-A   (1          )  00               I LIKE MATH (GRADE 4):  AGREE
002   LKMATH-U   (2          )  10               I LIKE MATH (GRADE 4):  UNDECIDED
003   LKMATH-D   (3,M        )  01               I LIKE MATH (GRADE 4):  DISAGREE, MISSING

CONDITIONING ID:              MATH0023
DESCRIPTION:                  AGREE/DISAGREE: I AM GOOD AT MATH (GRADE 4)
GRADES/ASSESSMENTS:           N04, S04
GROUP LABEL:                  GOODMAT4            LENGTH OF CONTRAST FIELD      :   2
NAEP ID:                      M811103             DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:             CLASS               NUMBER OF SPECIFICATION RECORDS:   3

001   GDMATH-A   (1          )  00               I AM GOOD AT MATH (GRADE 4): AGREE
002   GDMATH-U   (2          )  10               I AM GOOD AT MATH (GRADE 4): UNDECIDED
003   GDMATH-D   (3,M        )  01               I AM GOOD AT MATH (GRADE 4): DISAGREE, MISSING

CONDITIONING ID:              MATH0024
DESCRIPTION:                  AGREE/DISAGREE: UNDERSTAND MOST OF MATH CLASS (GRADE 4)
GRADES/ASSESSMENTS:           N04, S04
GROUP LABEL:                  USTDMAT4            LENGTH OF CONTRAST FIELD      :   2
NAEP ID:                      M811106             DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:             CLASS               NUMBER OF SPECIFICATION RECORDS:   3

001   USMATH-A   (1          )  00               UNDERSTAND MATH (GRADE 4): AGREE
002   USMATH-U   (2          )  10               UNDERSTAND MATH (GRADE 4): UNDECIDED
003   USMATH-D   (3,M        )  01               UNDERSTAND MATH (GRADE 4): DISAGREE, MISSING

CONDITIONING ID:              MATH0025
DESCRIPTION:                  AGREE/DISAGREE: MATH IS MORE FOR BOYS THAN FOR GIRLS (GRADE 4)
GRADES/ASSESSMENTS:           N04, S04
GROUP LABEL:                  MAT4BOY4            LENGTH OF CONTRAST FIELD      :   2
NAEP ID:                      M811104             DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:             CLASS               NUMBER OF SPECIFICATION RECORDS:   3

001   MA4BOY-A   (1,M        )  00               MATH FOR BOYS (GRADE 4): AGREE, MISSING
002   MA4BOY-U   (2          )  10               MATH FOR BOYS (GRADE 4): UNDECIDED
003   MA4BOY-D   (3          )  01               MATH FOR BOYS (GRADE 4): DISAGREE

CONDITIONING ID:              MATH0026
DESCRIPTION:                  AGREE/DISAGREE: MATH MOSTLY MEMORIZING FACTS (GRADE 4)
GRADES/ASSESSMENTS:           N04, S04
GROUP LABEL:                  MATMEMF4            LENGTH OF CONTRAST FIELD      :   2
NAEP ID:                      M811107             DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:             CLASS               NUMBER OF SPECIFICATION RECORDS:   3
```

```
001   MATMEM-A   (1            )  00          MATH IS MEMORIZING FACTS (GRADE 4): AGREE
002   MATMEM-U   (2            )  10          MATH IS MEMORIZING FACTS (GRADE 4): UNDECIDED
003   MATMEM-D   (3,M          )  01          MATH IS MEMORIZING FACTS (GRADE 4): DISAGREE, MISSING

CONDITIONING ID:          MATH0027
DESCRIPTION:              AGREE/DISAGREE: PEOPLE USE MATH IN JOBS (GRADE 4)
GRADES/ASSESSMENTS:       N04, S04
GROUP LABEL:              MATJOBS4            LENGTH OF CONTRAST FIELD       :  2
NAEP ID:                  M811102             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS               NUMBER OF SPECIFICATION RECORDS:  3

001   MATJOB-A   (1            )  00          USE MATH IN JOBS (GRADE 4): AGREE
002   MATJOB-U   (2            )  10          USE MATH IN JOBS (GRADE 4): UNDECIDED
003   MATJOB-D   (3,M          )  01          USE MATH IN JOBS (GRADE 4): DISAGREE, MISSING

CONDITIONING ID:          MATH0028
DESCRIPTION:              AGREE/DISAGREE: MATH USED FOR SOLVING EVERYDAY PROBLEMS (GRADE 4)
GRADES/ASSESSMENTS:       N04, S04
GROUP LABEL:              MAT4PRB4            LENGTH OF CONTRAST FIELD       :  2
NAEP ID:                  M811105             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS               NUMBER OF SPECIFICATION RECORDS:  3

001   MA4PRB-A   (1            )  00          USE MATH FOR SOLVING PROBLEMS (GRADE 4): AGREE
002   MA4PRB-U   (2            )  10          USE MATH FOR SOLVING PROBLEMS (GRADE 4): UNDECIDED
003   MA4PRB-D   (3,M          )  01          USE MATH FOR SOLVING PROBLEMS (GRADE 4): DISAGREE, MISSING

CONDITIONING ID:          MATH0029
DESCRIPTION:              AGREE/DISAGREE: IF CHOICE, WOULD NOT STUDY MORE MATH (GRADE 4)
GRADES/ASSESSMENTS:       N04, S04
GROUP LABEL:              NSTDMAT4            LENGTH OF CONTRAST FIELD       :  2
NAEP ID:                  M811108             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS               NUMBER OF SPECIFICATION RECORDS:  3

001   NSTDMA-A   (1            )  00          NO MORE MATH STUDY (GRADE 4): AGREE
002   NSTDMA-U   (2            )  10          NO MORE MATH STUDY (GRADE 4): UNDECIDED
003   NSTDMA-D   (3,M          )  01          NO MORE MATH STUDY (GRADE 4): DISAGREE, MISSING

CONDITIONING ID:          MATH0030
DESCRIPTION:              AGREE/DISAGREE: I LIKE MATH (GRADES 8, 12)
GRADES/ASSESSMENTS:       N08, S08, N12
GROUP LABEL:              LIKEMATH            LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                  M810701             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS               NUMBER OF SPECIFICATION RECORDS:  5

001   LIKMAT-SA  (1            )  0000        I LIKE MATH (GRADE 8, 12):  STRONGLY AGREE
002   LIKMAT-A   (2            )  1000        I LIKE MATH (GRADE 8, 12):  AGREE
003   LIKMAT-U   (3            )  0100        I LIKE MATH (GRADE 8, 12):  UNDECIDED
004   LIKMAT-D   (4            )  0010        I LIKE MATH (GRADE 8, 12):  DISAGREE
005   LIKMAT-SD  (5,M          )  0001        I LIKE MATH (GRADE 8, 12):  STRONGLY DISAGREE, MISSING

CONDITIONING ID:          MATH0031
DESCRIPTION:              AGREE/DISAGREE: I AM GOOD AT MATH (GRADES 8, 12)
GRADES/ASSESSMENTS:       N08, S08, N12
GROUP LABEL:              GOODMATH            LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                  M810703             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS               NUMBER OF SPECIFICATION RECORDS:  5

001   GDMAT-SA   (1            )  0000        I AM GOOD AT MATH (GRADE 8, 12): STRONGLY AGREE
002   GDMAT-A    (2            )  1000        I AM GOOD AT MATH (GRADE 8, 12): AGREE
003   GDMAT-U    (3            )  0100        I AM GOOD AT MATH (GRADE 8, 12): UNDECIDED
004   GDMAT-D    (4            )  0010        I AM GOOD AT MATH (GRADE 8, 12): DISAGREE
005   GDMAT-SD   (5,M          )  0001        I AM GOOD AT MATH (GRADE 8, 12): STRONGLY DISAGREE, MISSING

CONDITIONING ID:          MATH0032
DESCRIPTION:              AGREE/DISAGREE: IF CHOICE, WOULD NOT STUDY ANY MORE MATH (GRADE 8)
GRADES/ASSESSMENTS:       N08, S08
GROUP LABEL:              NSTDMATH            LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                  M810706             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS               NUMBER OF SPECIFICATION RECORDS:  5
```

```
001   NSMAT-SA   (1,M        )  0000        NO MORE MATH STUDY (GRADE 8): STRONGLY AGREE, MISSING
002   NSMAT-A    (2          )  1000        NO MORE MATH STUDY (GRADE 8): AGREE
003   NSMAT-U    (3          )  0100        NO MORE MATH STUDY (GRADE 8): UNDECIDED
004   NSMAT-D    (4          )  0010        NO MORE MATH STUDY (GRADE 8): DISAGREE
005   NSMAT-SD   (5          )  0001        NO MORE MATH STUDY (GRADE 8): STRONGLY DISAGREE

CONDITIONING ID:            MATH0033
DESCRIPTION:                AGREE/DISAGREE: I UNDERSTAND MOST OF MATH CLASS (GRADE 8, 12)
GRADES/ASSESSMENTS:         N08, S08, N12
GROUP LABEL:                UNDSTMAT            LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                    M810707             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:  5

001   USMAT-SA   (1          )  0000        UNDERSTAND MATH (GRADE 8, 12): STRONGLY AGREE
002   USMAT-A    (2          )  1000        UNDERSTAND MATH (GRADE 8, 12): AGREE
003   USMAT-U    (3          )  0100        UNDERSTAND MATH (GRADE 8, 12): UNDECIDED
004   USMAT-D    (4          )  0010        UNDERSTAND MATH (GRADE 8, 12): DISAGREE
005   USMAT-SD   (5,M        )  0001        UNDERSTAND MATH (GRADE 8, 12): STRONGLY DISAGREE, MISSING

CONDITIONING ID:            MATH0034
DESCRIPTION:                AGREE/DISAGREE: MATH IS MORE FOR BOYS THAN FOR GIRLS (GRADE 8, 12)
GRADES/ASSESSMENTS:         N08, S08, N12
GROUP LABEL:                MAT4BOYS            LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                    M810704             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:  5

001   M4BOY-SA   (1          )  0[illegible]0        MATH FOR BOYS (GRADE 8, 12): STRONGLY AGREE
002   M4BOY-A    (2          )  1000        MATH FOR BOYS (GRADE 8, 12): AGREE
003   M4BOY-U    (3          )  0100        MATH FOR BOYS (GRADE 8, 12): UNDECIDED
004   M4BOY-D    (4          )  0010        MATH FOR BOYS (GRADE 8, 12): DISAGREE
005   M4BOY-SD   (5,M        )  0001        MATH FOR BOYS (GRADE 8, 12): STRONGLY DISAGREE, MISSING

CONDITIONING ID:            MATH0035
DESCRIPTION:                AGREE/DISAGREE: MATH IS MOSTLY MEMORIZING FACTS (GRADE 8, 12)
GRADES/ASSESSMENTS:         N08, S08, N12
GROUP LABEL:                MATHMEMF            LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                    M810706             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:  5

001   MATMF-SA   (1          )  0000        MATH IS MEMORIZING FACTS (GR 8, 12): STRONGLY AGREE
002   MATMF-A    (2          )  1000        MATH IS MEMORIZING FACTS (GR 8, 12): AGREE
003   MATMF-U    (3          )  0100        MATH IS MEMORIZING FACTS (GR 8, 12): UNDECIDED
004   MATMF-D    (4          )  0010        MATH IS MEMORIZING FACTS (GR 8, 12): DISAGREE
005   MATMF-SD   (5,M        )  0001        MATH IS MEMORIZING FACTS (GR 8, 12): STRONGLY DISAG, MISSING

CONDITIONING ID:            MATH0036
DESCRIPTION:                AGREE/DISAGREE: ALMOST ALL PEOPLE USE MATH IN THEIR JOBS (GRADE 8, 12)
GRADES/ASSESSMENTS:         N08, S08, N12
GROUP LABEL:                MATHJOBS            LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                    M810702             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:  5

001   MAJOB-SA   (1          )  0000        USE MATH EVERYDAY PROBLEMS (GRADE 8, 12): STRONGLY AGREE
002   MAJOB-A    (2          )  1000        USE MATH IN JOBS (GRADE 8, 12): AGREE
003   MAJOB-U    (3          )  0100        USE MATH IN JOBS (GRADE 8, 12): UNDECIDED
004   MAJOB-D    (4          )  0010        USE MATH IN JOBS (GRADE 8, 12): DISAGREE
005   MAJOB-SD   (5,M        )  0001        USE MATH IN JOBS (GRADE 8, 12): STRONGLY DISAGREE, MISSING

CONDITIONING ID:            MATH0037
DESCRIPTION:                AGREE/DISAGREE: MATH IS USEFUL FOR SOLVING EVERYDAY PROBLEMS (GRADE 8, 12)
GRADES/ASSESSMENTS:         N08, S08, N12
GROUP LABEL:                MATHPROB            LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                    M810705             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:  5

001   MAPRB-SA   (1          )  0000        MATH FOR EVERYDAY PROBS (GR 8, 12): STRONGLY AGREE
002   MAPRB-A    (2          )  1000        MATH FOR EVERYDAY PROBS (GR 8, 12): AGREE
003   MAPRB-U    (3          )  0100        MATH FOR EVERYDAY PROBS (GR 8, 12): UNDECIDED
004   MAPRB-D    (4          )  0010        MATH FOR EVERYDAY PROBS (GR 8, 12): DISAGREE
```

```
005   MAPRB-SD   (5,M          )  0001        MATH FOR EVERYDAY PROBS (GR 8, 12): STRONGLY DISAGREE, MISSNG

CONDITIONING ID:               MATH0050
DESCRIPTION:                   ABOUT HOW MANY QUESTIONS DID YOU GET RIGHT?
GRADES/ASSESSMENTS:            N04, S04, N08, S08, N12
GROUP LABEL:                   #QUESTN+             LENGTH OF CONTRAST FIELD        :  3
NAEP ID:                       MM00101              DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:              CLASS                NUMBER OF SPECIFICATION RECORDS:   4

001   #QUEST+1   (1            )  000         NUMBER QUESTIONS RIGHT:  ALMOST ALL
002   #QUEST+2   (2            )  100         NUMBER QUESTIONS RIGHT:  MORE THAN HALF
003   #QUEST+3   (3            )  010         NUMBER QUESTIONS RIGHT:  ABOUT HALF
004   #QUEST+4   (4,M          )  001         NUMBER QUESTIONS RIGHT:  LESS THAN HALF, MISSING

CONDITIONING ID:               MATH0051
DESCRIPTION:                   HOW HARD WAS THIS TEST COMPARED TO OTHERS?
GRADES/ASSESSMENTS:            N04, S04, N08, S08, N12
GROUP LABEL:                   TEST_DIF             LENGTH OF CONTRAST FIELD        :  3
NAEP ID:                       MM00201              DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:              CLASS                NUMBER OF SPECIFICATION RECORDS:   4

001   TESTDIF1   (1,M          )  000         TEST DIFFICULTY:  MUCH HARDER THAN OTHERS
002   TESTDIF2   (2            )  100         TEST DIFFICULTY:  HARDER THAN OTHERS
003   TESTDIF3   (3            )  010         TEST DIFFICULTY:  ABOUT AS HARD AS OTHERS
004   TESTDIF4   (4            )  001         TEST DIFFICULTY:  EASIER THAN OTHERS

CONDITIONING ID:               MATH0052
DESCRIPTION:                   HOW HARD DID YOU TRY ON THIS TEST COMPRAED TO OTHERS?
GRADES/ASSESSMENTS:            N04, S04, N08, S08, N12
GROUP LABEL:                   TEST_EFF             LENGTH OF CONTRAST FIELD        :  3
NAEP ID:                       MM00301              DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:              CLASS                NUMBER OF SPECIFICATION RECORDS:   4

001   TESTEFF1   (1,M          )  000         TEST EFFORT:  MUCH HARDER THAN OTHERS
002   TESTEFF2   (2            )  100         TEST EFFORT:  HARDER THAN OTHERS
003   TESTEFF3   (3            )  010         TEST EFFORT:  ABOUT AS HARD AS OTHERS
004   TESTEFF4   (4            )  001         TEST EFFORT:  NOT AS HARD AS OTHERS

CONDITIONING ID:               MATH0053
DESCRIPTION:                   HOW IMPORTANT WAS IT TO YOU TO DO WELL?
GRADES/ASSESSMENTS:            N04, S04, N08, S08, N12
GROUP LABEL:                   TEST_IMP             LENGTH OF CONTRAST FIELD        :  4
NAEP ID:                       MM00401              DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:              CLASS                NUMBER OF SPECIFICATION RECORDS:   5

001   TESTIMP1   (1            )  0000        TEST IMPORTANCE:  VERY IMPORTANT
002   TESTIMP2   (2            )  1000        TEST IMPORTANCE:  IMPORTANT
003   TESTIMP3   (3            )  0100        TEST IMPORTANCE:  SOMEWHAT IMPORTANT
004   TESTIMP4   (4            )  0010        TEST IMPORTANCE:  NOT VERY IMPORTANT
005   TESTIMP7   (M            )  0001        TEST IMPORTANCE:  MISSING

CONDITIONING ID:               MATH0054
DESCRIPTION:                   HOW OFTEN WERE YOU ASKED TO PROVIDE DETAILED SOLUTIONS ON TESTS?
GRADES/ASSESSMENTS:            N04, S04, N08, S08, N12
GROUP LABEL:                   DSOLUTNS             LENGTH OF CONTRAST FIELD        :  3
NAEP ID:                       MM00501              DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:              CLASS                NUMBER OF SPECIFICATION RECORDS:   4

001   DSOLUTN1   (1            )  000         DETAILED SOLUTIONS:  AT LEAST ONCE A WEEK
002   DSOLUTN2   (2            )  100         DETAILED SOLUTIONS:  ONCE OR TWICE A MONTH
003   DSOLUTN3   (3            )  010         DETAILED SOLUTIONS:  ONCE OR TWICE A YEAR
004   DSOLUTN4   (4,M          )  001         DETAILED SOLUTIONS:  NEVER, MISSING

CONDITIONING ID:               SCHL0006
DESCRIPTION:                   DOES SCHOOL OFFER ALGEBRA TO 8TH GRADE FOR HS CREDIT?
GRADES/ASSESSMENTS:            N08, S08
GROUP LABEL:                   ALG4HSCR             LENGTH OF CONTRAST FIELD        :  2
NAEP ID:                       C034600              DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:              CLASS                NUMBER OF SPECIFICATION RECORDS:   3
```

```
001  ALGAHS-Y   (1          )  00        ALGEBRA FOR HS CREDIT:  YES
002  ALGAHS-N   (2          )  10        ALGEBRA FOR HS CREDIT:  NO
003  ALGAHS-?   (M          )  01        ALGEBRA FOR HS CREDIT:  MISSING

CONDITIONING ID:              SCHL0007
DESCRIPTION:                  WHO TEACHES ENGLISH/LANGUAGE ARTS TO 8TH GRADE?
GRADES/ASSESSMENTS:           N08, S08
GROUP LABEL:                  TSUB_ENG         LENGTH OF CONTRAST FIELD       :  2
NAEP ID:                      C034701          DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:             CLASS            NUMBER OF SPECIFICATION RECORDS:  3

001  ENG-T>1S   (1,M        )  00        8TH-GRADE ENGLISH:  TEACHERS WITH MORE THAN ONE SUBJECT
002  ENG-T=1S   (2          )  10        8TH-GRADE ENGLISH:  TEACHERS WITH ONE SUBJECT
003  ENG-SNT    (3          )  01        8TH-GRADE ENGLISH:  SUBJECT NOT TAUGHT

CONDITIONING ID:              SCHL0008
DESCRIPTION:                  WHO TEACHES MATHEMATICS TO 8TH GRADE?
GRADES/ASSESSMENTS:           N08, S08
GROUP LABEL:                  TSUB_MAT         LENGTH OF CONTRAST FIELD       :  2
NAEP ID:                      C034702          DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:             CLASS            NUMBER OF SPECIFICATION RECORDS:  3

001  MAT-T>1S   (1,M        )  00        8TH-GRADE MATH:  TEACHERS WITH MORE THAN ONE SUBJECT
002  MAT-T=1S   (2          )  10        8TH-GRADE MATH:  TEACHERS WITH ONE SUBJECT
003  MAT-SNT    (3          )  01        8TH-GRADE MATH:  SUBJECT NOT TAUGHT

CONDITIONING ID:              SCHL0009
DESCRIPTION:                  HAS READING BEEN IDENTIFIED AS A PRIORITY? (GRADE 4)
GRADES/ASSESSMENTS:           N04, S04, N08, S08
GROUP LABEL:                  PRIOR-RD         LENGTH OF CONTRAST FIELD       :  2
NAEP ID:                      C031601          DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:             CLASS            NUMBER OF SPECIFICATION RECORDS:  3

001  RPRIOR-Y   (1          )  00        READING PRIORITY:  YES
002  RPRIOR-N   (2          )  10        READING PRIORITY:  NO
003  RPRIOR-?   (M          )  01        READING PRIORITY:  MISSING

CONDITIONING ID:              SCHL0010
DESCRIPTION:                  HAS WRITING BEEN IDENTIFIED AS A PRIORITY? (GRADE 4)
GRADES/ASSESSMENTS:           N04, S04, N08, S08
GROUP LABEL:                  PRIOR-WR         LENGTH OF CONTRAST FIELD       :  2
NAEP ID:                      C031602          DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:             CLASS            NUMBER OF SPECIFICATION RECORDS:  3

001  WPRIOR-Y   (1          )  00        WRITING PRIORITY:  YES
002  WPRIOR-N   (2          )  10        WRITING PRIORITY:  NO
003  WPRIOR-?   (M          )  01        WRITING PRIORITY:  MISSING

CONDITIONING ID:              SCHL0011
DESCRIPTION:                  HAS MATHEMATICS BEEN IDENTIFIED AS A PRIORITY? (GRADE 4)
GRADES/ASSESSMENTS:           N04, S04, N08, S08
GROUP LABEL:                  PRIOR-MA         LENGTH OF CONTRAST FIELD       :  2
NAEP ID:                      C031603          DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:             CLASS            NUMBER OF SPECIFICATION RECORDS:  3

001  MPRIOR-Y   (1          )  00        MATH PRIORITY:  YES
002  MPRIOR-N   (2          )  10        MATH PRIORITY:  NO
003  MPRIOR-?   (M          )  01        MATH PRIORITY:  MISSING

CONDITIONING ID:              SCHL0012
DESCRIPTION:                  WHAT PERCENT OF STUDENTS RECEIVE SUBSIDIZED LUNCH?
GRADES/ASSESSMENTS:           N04, S04, N08, S08, N12
GROUP LABEL:                  %SUBLUN          LENGTH OF CONTRAST FIELD       :  8
NAEP ID:                      C032001          DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:             CLASS            NUMBER OF SPECIFICATION RECORDS:  9

001  %SUBLUN1   (1          )  00000000  PERCENT SUBSIDIZED LUNCH:  NONE
002  %SUBLUN2   (2          )  10000000  PERCENT SUBSIDIZED LUNCH:  1-5%
003  %SUBLUN3   (3          )  01000000  PERCENT SUBSIDIZED LUNCH:  6-10%
```

```
004   %SUBLUN4   (4            )  00100000    PERCENT SUBSIDIZED LUNCH:  11-25%
005   %SUBLUN5   (5            )  00010000    PERCENT SUBSIDIZED LUNCH:  26-50%
006   %SUBLUN6   (6            )  00001000    PERCENT SUBSIDIZED LUNCH:  51-75%
007   %SUBLUN7   (7            )  00000100    PERCENT SUBSIDIZED LUNCH:  76-90%
008   %SUBLUN8   (8            )  00000010    PERCENT SUBSIDIZED LUNCH:  90-100%
009   %SUBLUN?   (M            )  00000001    PERCENT SUBSIDIZED LUNCH:  MISSING

CONDITIONING ID:           SCHL0013
DESCRIPTION:               WHAT PERCENT OF STUDENTS RECEIVE REMEDIAL READING?
GRADES/ASSESSMENTS:        N04, S04, N08, S08, N12
GROUP LABEL:               %REMDL-R           LENGTH OF CONTRAST FIELD       :  8
NAEP ID:                   C032002            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:          CLASS              NUMBER OF SPECIFICATION RECORDS:  9

001   %REMRED1   (1            )  00000000    PERCENT REMEDIAL READING:  NONE
002   %REMRED2   (2            )  10000000    PERCENT REMEDIAL READING:  1-5%
003   %REMRED3   (3            )  01000000    PERCENT REMEDIAL READING:  6-10%
004   %REMRED4   (4            )  00100000    PERCENT REMEDIAL READING:  11-25%
005   %REMRED5   (5            )  00010000    PERCENT REMEDIAL READING:  26-50%
006   %REMRED6   (6            )  00001000    PERCENT REMEDIAL READING:  51-75%
007   %REMRED7   (7            )  00000100    PERCENT REMEDIAL READING:  76-90%
008   %REMRED8   (8            )  00000010    PERCENT REMEDIAL READING:  90-100%
009   %REMRED?   (M            )  00000001    PERCENT REMEDIAL READING:  MISSING

CONDITIONING ID:           SCHL0014
DESCRIPTION:               WHAT PERCENT OF STUDENTS RECEIVE REMEDIAL MATHEMATICS?
GRADES/ASSESSMENTS:        N04, S04, N08, S08, N12
GROUP LABEL:               %REMDL-M           LENGTH OF CONTRAST FIELD       :  8
NAEP ID:                   C032003            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:          CLASS              NUMBER OF SPECIFICATION RECORDS:  9

001   %REMMAT1   (1            )  00000000    PERCENT REMEDIAL MATH:  NONE
002   %REMMAT2   (2            )  10000000    PERCENT REMEDIAL MATH:  1-5%
003   %REMMAT3   (3            )  01000000    PERCENT REMEDIAL MATH:  6-10%
004   %REMMAT4   (4            )  00100000    PERCENT REMEDIAL MATH:  11-25%
005   %REMMAT5   (5            )  00010000    PERCENT REMEDIAL MATH:  26-50%
006   %REMMAT6   (6            )  00001000    PERCENT REMEDIAL MATH:  51-75%
007   %REMMAT7   (7            )  00000100    PERCENT REMEDIAL MATH:  76-90%
008   %REMMAT8   (8            )  00000010    PERCENT REMEDIAL MATH:  90-100%
009   %REMMAT?   (M            )  00000001    PERCENT REMEDIAL MATH:  MISSING

CONDITIONING ID:           SCHL0015
DESCRIPTION:               WHAT PERCENTAGE OF STUDENTS ARE ENROLLED AT BEGINNING AND END OF SCHOOL YEAR?
GRADES/ASSESSMENTS:        N04, S04, N08, S08, N12
GROUP LABEL:               %ENR/YR            LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                   C033700            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:          CLASS              NUMBER OF SPECIFICATION RECORDS:  5

001   %ENR/YR1   (1            )  0000        YEAR LONG ENROLLMENT:  98-100 PERCENT
002   %ENR/YR2   (2            )  1000        YEAR LONG ENROLLMENT:  95-97 PERCENT
003   %ENR/YR3   (3            )  0100        YEAR LONG ENROLLMENT:  90-94 PERCENT
004   %ENR/YR4   (4            )  0010        YEAR LONG ENROLLMENT:  LESS THAN 90 PERCENT
005   %ENR/YR?   (M            )  0001        YEAR LONG ENROLLMENT:  MISSING

CONDITIONING ID:           SCHL0016
DESCRIPTION:               WHAT PERCENTAGE OF 4TH GRADERS RETAINED IN 91-92?
GRADES/ASSESSMENTS:        N04, S04, N08, S08
GROUP LABEL:               %4RETAIN           LENGTH OF CONTRAST FIELD       :  5
NAEP ID:                   C033800            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:          CLASS              NUMBER OF SPECIFICATION RECORDS:  6

001   %RETAIN1   (1            )  00000       % 4TH GRADE RETAINED:  0 PERCENT
002   %RETAIN2   (2            )  10000       % 4TH GRADE RETAINED:  1-2 PERCENT
003   %RETAIN3   (3            )  01000       % 4TH GRADE RETAINED:  3-5 PERCENT
004   %RETAIN4   (4            )  00100       % 4TH GRADE RETAINED:  6-10 PERCENT
005   %RETAIN5   (5            )  00010       % 4TH GRADE RETAINED:  MORE THAN 10 PERCENT
006   %RETAIN?   (M            )  00001       % 4TH GRADE RETAINED:  MISSING

CONDITIONING ID:           SCHL0017
```

```
DESCRIPTION:              WHAT PERCENTAGE OF TEACHERS NOT AT SCHOOL AT END OF YEAR?
GRADES/ASSESSMENTS:       N04, S04, N08, S08, N12
GROUP LABEL:              %T_LEAVE              LENGTH OF CONTRAST FIELD       :   5
NAEP ID:                  C033903               DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:         CLASS                 NUMBER OF SPECIFICATION RECORDS:   6

001   %TLEAVE1   (1          )   00000          % TEACHERS LEAVE/YEAR: 0 PERCENT
002   %TLEAVE2   (2          )   10000          % TEACHERS LEAVE/YEAR: 1-2 PERCENT
003   %TLEAVE3   (3          )   01000          % TEACHERS LEAVE/YEAR: 3-5 PERCENT
004   %TLEAVE4   (4          )   00100          % TEACHERS LEAVE/YEAR: 6-10 PERCENT
005   %TLEAVE5   (5          )   00010          % TEACHERS LEAVE/YEAR: MORE THAN 10 PERCENT
006   %TLEAVE?   (M          )   00001          % TEACHERS LEAVE/YEAR: MISSING

CONDITIONING ID:          TCHR0001
DESCRIPTION:              HOW WELL DOES SCHOOL PROVIDE RESOURCES
GRADES/ASSESSMENTS:       N04, S04, N08, S08
GROUP LABEL:              RESOURCE              LENGTH OF CONTRAST FIELD       :   4
NAEP ID:                  T041201               DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:         CLASS                 NUMBER OF SPECIFICATION RECORDS:   5

001   RESOURC1   (1          )   0000           RESOURCES:  GET ALL
002   RESOURC2   (2          )   1000           RESOURCES:  GET MOST
003   RESOURC3   (3          )   0100           RESOURCES:  GET SOME
004   RESOURC4   (4          )   0010           RESOURCES:  DON'T GET
005   RESOURC?   (M,DNA      )   0001           RESOURCES:  MISSING, DOES NOT APPLY

CONDITIONING ID:          TCHR0002
DESCRIPTION:              TEACHER MATCH STATUS WITH STUDENT
GRADES/ASSESSMENTS:       N04, S04, N08, S08
GROUP LABEL:              T_MATCH               LENGTH OF CONTRAST FIELD       :   2
NAEP ID:                  TCHMTCH               DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:         CLASS                 NUMBER OF SPECIFICATION RECORDS:   3

001   TMCH-NO    (1,M        )   00             TEACHER MATCH:  NO MATCH
002   TMCH-PAR   (2          )   10             TEACHER MATCH:  PARTIAL MATCH
003   TMCH-COM   (3          )   01             TEACHER MATCH:  COMPLETE MATCH

CONDITIONING ID:          TMAT0001
DESCRIPTION:              WHAT IS THE MATH ABILITY OF STUDENTS IN THIS CLASS?
GRADES/ASSESSMENTS:       N04, S04, N08, S08
GROUP LABEL:              ABIL_MAT              LENGTH OF CONTRAST FIELD       :   4
NAEP ID:                  T044100               DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:         CLASS                 NUMBER OF SPECIFICATION RECORDS:   5

001   AB_MATH1   (1          )   0000           MATH ABILITY:  PRIMARILY HIGH ABILITY
002   AB_MATH2   (2          )   1000           MATH ABILITY:  PRIMARILY LOW ABILITY
003   AB_MATH3   (3          )   0100           MATH ABILITY:  PRIMARILY AVERAGE ABILITY
004   AB_MATH4   (4          )   0010           MATH ABILITY:  WIDELY MIXED ABILITY
005   AB_MATH?   (M,DNA      )   0001           MATH ABILITY:  MISSING, DOES NOT APPLY

CONDITIONING ID:          TMAT0002
DESCRIPTION:              AMOUNT OF MATH HOMEWORK ASSIGNED PER DAY (TEACHER)
GRADES/ASSESSMENTS:       N04, S04, N08, S08
GROUP LABEL:              T_MATHHW              LENGTH OF CONTRAST FIELD       :   6
NAEP ID:                  T044400               DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:         CLASS                 NUMBER OF SPECIFICATION RECORDS:   7

001   T_MATHW1   (1          )   000000         AMOUNT MATH HOMEWORK (TEACHER):  NONE
002   T_MATHW2   (2          )   100000         AMOUNT MATH HOMEWORK (TEACHER):  15 MINUTES
003   T_MATHW3   (3          )   010000         AMOUNT MATH HOMEWORK (TEACHER):  30 MINUTES
004   T_MATHW4   (4          )   001000         AMOUNT MATH HOMEWORK (TEACHER):  45 MINUTES
005   T_MATHW5   (5          )   000100         AMOUNT MATH HOMEWORK (TEACHER):  1 HOUR
006   T_MATHW6   (6          )   000010         AMOUNT MATH HOMEWORK (TEACHER):  MORE THAN 1 HOUR
007   T_MATHW?   (M,DNA      )   000001         AMOUNT MATH HOMEWORK (TEACHER):  MISSING, DOES NOT APPLY

CONDITIONING ID:          TMAT0003
DESCRIPTION:              HOW OFTEN DO STUDENTS DO MATH FROM TEXTBOOKS?
GRADES/ASSESSMENTS:       N04, S04, N08, S08
GROUP LABEL:              T_TXTBKS              LENGTH OF CONTRAST FIELD       :   4
```

```
NAEP ID:                  T044501          DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:         CLASS            NUMBER OF SPECIFICATION RECORDS:   5

001   T_TXTBK1   (1          )  0000       MATH FROM TEXTBOOKS (TEACHER):  ALMOST EVERY DAY
002   T_TXTBK2   (2          )  1000       MATH FROM TEXTBOOKS (TEACHER):  ONCE OR TWICE A WEEK
003   T_TXTBK3   (3          )  0100       MATH FROM TEXTBOOKS (TEACHER):  ONCE OR TWICE A MONTH
004   T_TXTBK4   (4          )  0010       MATH FROM TEXTBOOKS (TEACHER):  NEVER OR HARDLY EVER
005   T_TXTBK7   (M,DNA      )  0001       MATH FROM TEXTBOOKS (TEACHER):  MISSING, DOES NOT APPLY

CONDITIONING ID:          TMAT0004
DESCRIPTION:              HOW OFTEN DO STUDENTS DO MATH FROM WORKSHEETS?
GRADES/ASSESSMENTS:       N04, S04, N08, S08
GROUP LABEL:              T_WRKSHS         LENGTH OF CONTRAST FIELD       :   4
NAEP ID:                  T044502          DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:         CLASS            NUMBER OF SPECIFICATION RECORDS:   5

001   T_WRKSH1   (1          )  0000       MATH FROM WORKSHEETS (TEACHER): ALMOST EVERY DAY
002   T_WRKSH2   (2          )  1000       MATH FROM WORKSHEETS (TEACHER): ONCE OR TWICE A WEEK
003   T_WRKSH3   (3          )  0100       MATH FROM WORKSHEETS (TEACHER): ONCE OR TWICE A MONTH
004   T_WRKSH4   (4          )  0010       MATH FROM WORKSHEETS (TEACHER): NEVER OR HARDLY EVER
005   T_WRKSH7   (M,DNA      )  0001       MATH FROM WORKSHEETS (TEACHER): MISSING, DOES NOT APPLY

CONDITIONING ID:          TMAT0005
DESCRIPTION:              HOW OFTEN DO STUDENTS DO MATH IN SMALL GROUPS?
GRADES/ASSESSMENTS:       N04, S04, N08, S08
GROUP LABEL:              T_SMGRPS         LENGTH OF CONTRAST FIELD       :   4
NAEP ID:                  T044503          DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:         CLASS            NUMBER OF SPECIFICATION RECORDS:   5

001   T_SMGRP1   (1          )  0000       MATH IN SMALL GROUPS (TEACHER): ALMOST EVERY DAY
002   T_SMGRP2   (2          )  1000       MATH IN SMALL GROUPS (TEACHER): ONCE OR TWICE A WEEK
003   T_SMGRP3   (3          )  0100       MATH IN SMALL GROUPS (TEACHER): ONCE OR TWICE A MONTH
004   T_SMGRP4   (4          )  0010       MATH IN SMALL GROUPS (TEACHER): NEVER OR HARDLY EVER
005   T_SMGRP7   (M,DNA      )  0001       MATH IN SMALL GROUPS (TEACHER): MISSING, DOES NOT APPLY

CONDITIONING ID:          TMAT0006
DESCRIPTION:              HOW OFTEN DO STUDENTS WORK WITH OBJECTS?
GRADES/ASSESSMENTS:       N04, S04
GROUP LABEL:              T_OBJCTS         LENGTH OF CONTRAST FIELD       :   4
NAEP ID:                  T044504          DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:         CLASS            NUMBER OF SPECIFICATION RECORDS:   5

001   T_OBJCT1   (1          )  0000       WORK WITH OBJECTS (TEACHER): ALMOST EVERY DAY
002   T_OBJCT2   (2          )  1000       WORK WITH OBJECTS (TEACHER): ONCE OR TWICE A WEEK
003   T_OBJCT3   (3          )  0100       WORK WITH OBJECTS (TEACHER): ONCE OR TWICE A MONTH
004   T_OBJCT4   (4          )  0010       WORK WITH OBJECTS (TEACHER): NEVER OR HARDLY EVER
005   T_OBJCT7   (M,DNA      )  0001       WORK WITH OBJECTS (TEACHER): MISSING, DOES NOT APPLY

CONDITIONING ID:          TMAT0007
DESCRIPTION:              HOW OFTEN DO STUDENTS WORK WITH MEASUREMENT INSTR/GEOM SOLIDS (TEACHER)?
GRADES/ASSESSMENTS:       N08, S08
GROUP LABEL:              T_MI&GS          LENGTH OF CONTRAST FIELD       :   4
NAEP ID:                  T044512          DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:         CLASS            NUMBER OF SPECIFICATION RECORDS:   5

001   T_MI&GS1   (1          )  0000       MEASUREMENT INSTR/GEOM SOLIDS (TEACHER): ALMOST EVERY DAY
002   T_MI&GS2   (2          )  1000       MEASUREMENT INSTR/GEOM SOLIDS (TEACHER): ONCE OR TWICE A WEEK
003   T_MI&GS3   (3          )  0100       MEASUREMENT INSTR/GEOM SOLIDS (TEACHER): ONCE OR TWICE MONTH
004   T_MI&GS4   (4          )  0010       MEASUREMENT INSTR/GEOM SOLIDS (TEACHER): NEVER OR HARDLY EVER
005   T_MI&GS7   (M,DNA      )  0001       MEASUREMENT INSTR/GEOM SOLIDS (TEACHER): MISSNG, DOESNT APPLY

CONDITIONING ID:          TMAT0008
DESCRIPTION:              HOW OFTEN DO STUDENTS USE A CALCULATOR (TEACHER)?
GRADES/ASSESSMENTS:       N04, S04, N08, S08
GROUP LABEL:              T_CALCTR         LENGTH OF CONTRAST FIELD       :   4
NAEP ID:                  T044505          DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:         CLASS            NUMBER OF SPECIFICATION RECORDS:   5

001   T_CALCR1   (1          )  0000       USE CALCULATOR (TEACHER): ALMOST EVERY DAY
```

```
002   T_CALCR2   (2          )  1000        USE CALCULATOR (TEACHER): ONCE OR TWICE A WEEK
003   T_CALCR3   (3          )  0100        USE CALCULATOR (TEACHER): ONCE OR TWICE A MONTH
004   T_CALCR4   (4          )  0010        USE CALCULATOR (TEACHER): NEVER OR HARDLY EVER
005   T_CALCR?   (M,DNA      )  0001        USE CALCULATOR (TEACHER): MISSING, DOES NOT APPLY

CONDITIONING ID:            TMAT0009
DESCRIPTION:                HOW OFTEN DO STUDENTS USE A COMPUTER (TEACHER)?
GRADES/ASSESSMENTS:         N04, S04, N08, S08
GROUP LABEL:                T_COMPTR             LENGTH OF CONTRAST FIELD       :   4
NAEP ID:                    T044506              DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:           CLASS                NUMBER OF SPECIFICATION RECORDS:   5

001   T_CMPTR1   (1          )  0000        USE COMPUTER (TEACHER): ALMOST EVERY DAY
002   T_CMPTR2   (2          )  1000        USE COMPUTER (TEACHER): ONCE OR TWICE A WEEK
003   T_CMPTR3   (3          )  0100        USE COMPUTER (TEACHER): ONCE OR TWICE A MONTH
004   T_CMPTR4   (4          )  0010        USE COMPUTER (TEACHER): NEVER OR HARDLY EVER
005   T_CMPTR?   (M,DNA      )  0001        USE COMPUTER (TEACHER): MISSING, DOES NOT APPLY

CONDITIONING ID:            TMAT0010
DESCRIPTION:                HOW OFTEN DO STUDENTS WRITE ABOUT PROBLEM SOLVING (TEACHER)?
GRADES/ASSESSMENTS:         N04, S04, N08, S08
GROUP LABEL:                T_PRBSOL             LENGTH OF CONTRAST FIELD       :   4
NAEP ID:                    T044507              DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:           CLASS                NUMBER OF SPECIFICATION RECORDS:   5

001   T_PRBSL1   (1          )  0000        PROBLEM SOLVING (TEACHER): ALMOST EVERY DAY
002   T_PRBSL2   (2          )  1000        PROBLEM SOLVING (TEACHER): ONCE OR TWICE A WEEK
003   T_PRBSL3   (3          )  0100        PROBLEM SOLVING (TEACHER): ONCE OR TWICE A MONTH
004   T_PRBSL4   (4          )  0010        PROBLEM SOLVING (TEACHER): NEVER OR HARDLY EVER
005   T_PRBSL?   (M,DNA      )  0001        PROBLEM SOLVING (TEACHER): MISSING, DOES NOT APPLY

CONDITIONING ID:            TMAT0011
DESCRIPTION:                HOW OFTEN DO STUDENTS WRITE REPORTS/DO PROJECTS (TEACHER)?
GRADES/ASSESSMENTS:         N04, S04, N08, S08
GROUP LABEL:                T_REPPRJ             LENGTH OF CONTRAST FIELD       :   4
NAEP ID:                    T044508              DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:           CLASS                NUMBER OF SPECIFICATION RECORDS:   5

001   T_REPPJ1   (1          )  0000        REPORTS/PROJECTS (TEACHER): ALMOST EVERY DAY
002   T_REPPJ2   (2          )  1000        REPORTS/PROJECTS (TEACHER): ONCE OR TWICE A WEEK
003   T_REPPJ3   (3          )  0100        REPORTS/PROJECTS (TEACHER): ONCE OR TWICE A MONTH
004   T_REPPJ4   (4          )  0010        REPORTS/PROJECTS (TEACHER): NEVER OR HARDLY EVER
005   T_REPPJ?   (M,DNA      )  0001        REPORTS/PROJECTS (TEACHER): MISSING, DOES NOT APPLY

CONDITIONING ID:            TMAT0012
DESCRIPTION:                HOW OFTEN DO STUDENTS DISCUSS MATH WITH OTHER STUDENTS (TEACHER)?
GRADES/ASSESSMENTS:         N04, S04, N08, S08
GROUP LABEL:                T_DISMAT             LENGTH OF CONTRAST FIELD       :   4
NAEP ID:                    T044509              DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:           CLASS                NUMBER OF SPECIFICATION RECORDS:   5

001   T_DISMA1   (1          )  0000        STUDENTS DISCUSS MATH (TEACHER): ALMOST EVERY DAY
002   T_DISMA2   (2          )  1000        STUDENTS DISCUSS MATH (TEACHER): ONCE OR TWICE A WEEK
003   T_DISMA3   (3          )  0100        STUDENTS DISCUSS MATH (TEACHER): ONCE OR TWICE A MONTH
004   T_DISMA4   (4          )  0010        STUDENTS DISCUSS MATH (TEACHER): NEVER OR HARDLY EVER
005   T_DISMA?   (M,DNA      )  0001        STUDENTS DISCUSS MATH (TEACHER): MISSING, DOES NOT APPLY

CONDITIONING ID:            TMAT0013
DESCRIPTION:                HOW OFTEN DO STUDENTS WORK REAL LIFE PROBLEMS (TEACHER)?
GRADES/ASSESSMENTS:         N04, S04, N08, S08
GROUP LABEL:                T_RLPROB             LENGTH OF CONTRAST FIELD       :   4
NAEP ID:                    T044510              DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:           CLASS                NUMBER OF SPECIFICATION RECORDS:   5

001   T_RLPRB1   (1          )  0000        REAL LIFE PROBLEMS (TEACHER): ALMOST EVERY DAY
002   T_RLPRB2   (2          )  1000        REAL LIFE PROBLEMS (TEACHER): ONCE OR TWICE A WEEK
003   T_RLPRB3   (3          )  0100        REAL LIFE PROBLEMS (TEACHER): ONCE OR TWICE A MONTH
004   T_RLPRB4   (4          )  0010        REAL LIFE PROBLEMS (TEACHER): NEVER OR HARDLY EVER
005   T_RLPRB?   (M,DNA      )  0001        REAL LIFE PROBLEMS (TEACHER): MISSING, DOES NOT APPLY
```

```
CONDITIONING ID:            TMAT0014
DESCRIPTION:                HOW OFTEN DO STUDENTS MAKE UP MATH PROBLEMS (TEACHER)?
GRADES/ASSESSMENTS:         N04, S04, N08, S08
GROUP LABEL:                T_MUPROB          LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                    T044511           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS             NUMBER OF SPECIFICATION RECORDS:  5

001   T_MUPRB1   (1         )  0000           STUDENTS MAKE UP PROBLEMS (TEACHER): ALMOST EVERY DAY
002   T_MUPRB2   (2         )  1000           STUDENTS MAKE UP PROBLEMS (TEACHER): ONCE OR TWICE A WEEK
003   T_MUPRB3   (3         )  0100           STUDENTS MAKE UP PROBLEMS (TEACHER): ONCE OR TWICE A MONTH
004   T_MUPRB4   (4         )  0010           STUDENTS MAKE UP PROBLEMS (TEACHER): NEVER OR HARDLY EVER
005   T_MUPRB?   (M,DNA     )  0001           STUDENTS MAKE UP PROBLEMS (TEACHER): MISSING, DOES NOT APPLY

CONDITIONING ID:            TMAT0015
DESCRIPTION:                HOW MUCH EMPHASIS ON NUMBERS AND OPERATIONS?
GRADES/ASSESSMENTS:         N04, S04, N08, S08
GROUP LABEL:                EMP_N&OP          LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                    T044601           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS             NUMBER OF SPECIFICATION RECORDS:  4

001   EMP_N&O1   (1         )  000            EMPHASIS NUMBERS AND OPERATIONS:  HEAVY EMPHASIS
002   EMP_N&O2   (2         )  100            EMPHASIS NUMBERS AND OPERATIONS:  MODERATE EMPHASIS
003   EMP_N&O3   (3         )  010            EMPHASIS NUMBERS AND OPERATIONS:  LITTLE OR NONE
004   EMP_N&O?   (M,DNA     )  001            EMPHASIS NUMBERS AND OPERATIONS:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TMAT0016
DESCRIPTION:                HOW MUCH EMPHASIS ON MEASUREMENT?
GRADES/ASSESSMENTS:         N04, S04, N08, S08
GROUP LABEL:                EMP_MEAS          LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                    T044602           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS             NUMBER OF SPECIFICATION RECORDS:  4

001   EMP_MEA1   (1         )  000            EMPHASIS MEASUREMENT:  HEAVY EMPHASIS
002   EMP_MEA2   (2         )  100            EMPHASIS MEASUREMENT:  MODERATE EMPHASIS
003   EMP_MEA3   (3         )  010            EMPHASIS MEASUREMENT:  LITTLE OR NONE
004   EMP_MEA?   (M,DNA     )  001            EMPHASIS MEASUREMENT:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TMAT0017
DESCRIPTION:                HOW MUCH EMPHASIS ON GEOMETRY?
GRADES/ASSESSMENTS:         N04, S04, N08, S08
GROUP LABEL:                EMP_GEOM          LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                    T044603           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS             NUMBER OF SPECIFICATION RECORDS:  4

001   EMP_GEO1   (1         )  000            EMPHASIS GEOMETRY:  HEAVY EMPHASIS
002   EMP_GEO2   (2         )  100            EMPHASIS GEOMETRY:  MODERATE EMPHASIS
003   EMP_GEO3   (3         )  010            EMPHASIS GEOMETRY:  LITTLE OR NONE
004   EMP_GEO?   (M,DNA     )  001            EMPHASIS GEOMETRY:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TMAT0018
DESCRIPTION:                HOW MUCH EMPHASIS ON DATA ANALYSIS/STATISTICS/PROBABILITY? (GRADE 4)
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                EMP_DSP4          LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                    T044611           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS             NUMBER OF SPECIFICATION RECORDS:  4

001   EMP_DSP1   (1         )  000            EMPHASIS DATA ANALYSIS/STAT/PROB:  HEAVY EMPHASIS
002   EMP_DSP2   (2         )  100            EMPHASIS DATA ANALYSIS/STAT/PROB:  MODERATE EMPHASIS
003   EMP_DSP3   (3         )  010            EMPHASIS DATA ANALYSIS/STAT/PROB:  LITTLE OR NONE
004   EMP_DSP?   (M,DNA     )  001            EMPHASIS DATA ANALYSIS/STAT/PROB:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TMAT0019
DESCRIPTION:                HOW MUCH EMPHASIS ON DATA ANALYSIS/STATISTICS/PROBABILITY? (GRADE 8)
GRADES/ASSESSMENTS:         N08, S08
GROUP LABEL:                EMP_DSP8          LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                    T044604           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS             NUMBER OF SPECIFICATION RECORDS:  4

001   EMP_DSP1   (1         )  000            EMPHASIS DATA ANALYSIS/STAT/PROB:  HEAVY EMPHASIS
```

BEST COPY AVAILABLE

```
002   EMP_DSP2   (2          )  100        EMPHASIS DATA ANALYSIS/STAT/PROB:  MODERATE EMPHASIS
003   EMP_DSP3   (3          )  010        EMPHASIS DATA ANALYSIS/STAT/PROB:  LITTLE OR NONE
004   EMP_DSP?   (M,DNA      )  001        EMPHASIS DATA ANALYSIS/STAT/PROB:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TMAT0020
DESCRIPTION:                HOW MUCH EMPHASIS ON ALGEBRA AND FUNCTIONS? (GRADE 4)
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                EMP_ALG4           LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                    T044612            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS              NUMBER OF SPECIFICATION RECORDS:  4

001   EMP_ALG1   (1          )  000        EMPHASIS ALGEBRA AND FUNCTIONS:  HEAVY EMPHASIS
002   EMP_ALG2   (2          )  100        EMPHASIS ALGEBRA AND FUNCTIONS:  MODERATE EMPHASIS
003   EMP_ALG3   (3          )  010        EMPHASIS ALGEBRA AND FUNCTIONS:  LITTLE OR NONE
004   EMP_ALG?   (M,DNA      )  001        EMPHASIS ALGEBRA AND FUNCTIONS:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TMAT0021
DESCRIPTION:                HOW MUCH EMPHASIS ON ALGEBRA AND FUNCTIONS? (GRADE 8)
GRADES/ASSESSMENTS:         N08, S08
GROUP LABEL:                EMP_ALG8           LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                    T044605            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS              NUMBER OF SPECIFICATION RECORDS:  4

001   EMP_ALG1   (1          )  000        EMPHASIS ALGEBRA AND FUNCTIONS:  HEAVY EMPHASIS
002   EMP_ALG2   (2          )  100        EMPHASIS ALGEBRA AND FUNCTIONS:  MODERATE EMPHASIS
003   EMP_ALG3   (3          )  010        EMPHASIS ALGEBRA AND FUNCTIONS:  LITTLE OR NONE
004   EMP_ALG?   (M,DNA      )  001        EMPHASIS ALGEBRA AND FUNCTIONS:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TMAT0022
DESCRIPTION:                HOW MUCH EMPHASIS ON LEARNING FACTS/CONCEPTS?
GRADES/ASSESSMENTS:         N04, S04, N08, S08
GROUP LABEL:                EMP_F/C            LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                    T044606            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS              NUMBER OF SPECIFICATION RECORDS:  4

001   EMP_F/C1   (1          )  000        EMPHASIS FACTS/CONCEPTS:  HEAVY EMPHASIS
002   EMP_F/C2   (2          )  100        EMPHASIS FACTS/CONCEPTS:  MODERATE EMPHASIS
003   EMP_F/C3   (3          )  010        EMPHASIS FACTS/CONCEPTS:  LITTLE OR NONE
004   EMP_F/C?   (M,DNA      )  001        EMPHASIS FACTS/CONCEPTS:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TMAT0023
DESCRIPTION:                HOW MUCH EMPHASIS ON LEARNING SKILLS/PROCEDURES?
GRADES/ASSESSMENTS:         N04, S04, N08, S08
GROUP LABEL:                EMP_S/P            LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                    T044607            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS              NUMBER OF SPECIFICATION RECORDS:  4

001   EMP_S/P1   (1          )  000        EMPHASIS SKILLS/PROCEDURES: HEAVY EMPHASIS
002   EMP_S/P2   (2          )  100        EMPHASIS SKILLS/PROCEDURES: MODERATE EMPHASIS
003   EMP_S/P3   (3          )  010        EMPHASIS SKILLS/PROCEDURES: LITTLE OR NONE
004   EMP_S/P?   (M,DNA      )  001        EMPHASIS SKILLS/PROCEDURES: MISSING, DOES NOT APPLY

CONDITIONING ID:            TMAT0024
DESCRIPTION:                HOW MUCH EMPHASIS ON REASONING/ANALYSIS?
GRADES/ASSESSMENTS:         N04, S04, N08, S08
GROUP LABEL:                EMP_R/A            LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                    T044608            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS              NUMBER OF SPECIFICATION RECORDS:  4

001   EMP_R/A1   (1          )  000        EMPHASIS REASONING/ANALYSIS: HEAVY EMPHASIS
002   EMP_R/A2   (2          )  100        EMPHASIS REASONING/ANALYSIS: MODERATE EMPHASIS
003   EMP_R/A3   (3          )  010        EMPHASIS REASONING/ANALYSIS: LITTLE OR NONE
004   EMP_R/A?   (M,DNA      )  001        EMPHASIS REASONING/ANALYSIS: MISSING, DOES NOT APPLY

CONDITIONING ID:            TMAT0025
DESCRIPTION:                HOW MUCH EMPHASIS ON COMMUNICATING MATH IDEAS?
GRADES/ASSESSMENTS:         N04, S04, N08, S08
GROUP LABEL:                EMP_CMI            LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                    T044609            DEGREES OF FREEDOM PER CONTRAST:  1
```

```
TYPE OF CONTRAST:          CLASS               NUMBER OF SPECIFICATION RECORDS:   4

001   EMP_CMI1   (1           )  000        EMPHASIS COMMUNICATING MATH IDEAS: HEAVY EMPHASIS
002   EMP_CMI2   (2           )  100        EMPHASIS COMMUNICATING MATH IDEAS: MODERATE EMPHASIS
003   EMP_CMI3   (3           )  010        EMPHASIS COMMUNICATING MATH IDEAS: LITTLE OR NONE
004   EMP_CMI?   (M,DNA       )  001        EMPHASIS COMMUNICATING MATH IDEAS: MISSING, DOES NOT APPLY

CONDITIONING ID:           TMAT0026
DESCRIPTION:               HOW MUCH EMPHASIS ON APPRECIATING MATHEMATICS?
GRADES/ASSESSMENTS:        N04, S04, N08, S08
GROUP LABEL:               EMP_MAP             LENGTH OF CONTRAST FIELD       :   3
NAEP ID:                   T044610             DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:          CLASS               NUMBER OF SPECIFICATION RECORDS:   4

001   EMP_MAP1   (1           )  000        EMPHASIS MATH APPRECIATION: HEAVY EMPHASIS
002   EMP_MAP2   (2           )  100        EMPHASIS MATH APPRECIATION: MODERATE EMPHASIS
003   EMP_MAP3   (3           )  010        EMPHASIS MATH APPRECIATION: LITTLE OR NONE
004   EMP_MAP?   (M,DNA       )  001        EMPHASIS MATH APPRECIATION: MISSING, DOES NOT APPLY

CONDITIONING ID:           TMAT0027
DESCRIPTION:               DO YOU PERMIT UNRESTRICTED USE OF CALCULATORS? (TEACHER)
GRADES/ASSESSMENTS:        N04, S04, N08, S08
GROUP LABEL:               CALC_UNR            LENGTH OF CONTRAST FIELD       :   2
NAEP ID:                   T045401             DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:          CLASS               NUMBER OF SPECIFICATION RECORDS:   3

001   CALUNR-Y   (1           )  00         UNRESTRICTED CALCULATOR USE:  YES
002   CALUNR-N   (2           )  10         UNRESTRICTED CALCULATOR USE:  NO
003   CALUNR-?   (M,DNA       )  01         UNRESTRICTED CALCULATOR USE:  MISSING, DOES NOT APPLY

CONDITIONING ID:           TMAT0028
DESCRIPTION:               DO YOU PERMIT USE OF CALCULATORS ON TESTS? (TEACHER)
GRADES/ASSESSMENTS:        N04, S04, N08, S08
GROUP LABEL:               CALC_TST            LENGTH OF CONTRAST FIELD       :   2
NAEP ID:                   T044801             DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:          CLASS               NUMBER OF SPECIFICATION RECORDS:   3

001   CALTST-Y   (1           )  00         CALCULATOR USE ON TESTS:  YES
002   CALTST-N   (2           )  10         CALCULATOR USE ON TESTS:  NO
003   CALTST-?   (M,DNA       )  01         CALCULATOR USE ON TESTS:  MISSING, DOES NOT APPLY

CONDITIONING ID:           TMAT0029
DESCRIPTION:               TEACHER HOURS SPENT IN IN-SERVICE MATHEMATICS EDUCATION
GRADES/ASSESSMENTS:        N04, S04, N08, S08
GROUP LABEL:               INSERV_M            LENGTH OF CONTRAST FIELD       :   5
NAEP ID:                   T040901             DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:          CLASS               NUMBER OF SPECIFICATION RECORDS:   6

001   INSERVM1   (1           )  00000      MATH HOURS IN-SERVICE:  NONE
002   INSERVM2   (2           )  10000      MATH HOURS IN-SERVICE:  LESS THAN 6 HOURS
003   INSERVM3   (3           )  01000      MATH HOURS IN-SERVICE:  6-15 HOURS
004   INSERVM4   (4           )  00100      MATH HOURS IN-SERVICE:  16-35 HOURS
005   INSERVM5   (5           )  00010      MATH HOURS IN-SERVICE:  MORE THAN 35 HOURS
006   INSERVM?   (M,DNA       )  00001      MATH HOURS IN-SERVICE:  MISSING, DOES NOT APPLY

CONDITIONING ID:           TMAT0030
DESCRIPTION:               ONE OR MORE COLLEGE COURSES IN SEVEN SUBJECTS
GRADES/ASSESSMENTS:        N04, S04, N08, S08
GROUP LABEL:               T_MATHCR            LENGTH OF CONTRAST FIELD       :   2
NAEP ID:                   TMATHCR             DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:          CLASS               NUMBER OF SPECIFICATION RECORDS:   3

001   T_MATCR1   (1,M         )  00         TEACHER COURSES:  0-3
002   T_MATCR2   (2           )  10         TEACHER COURSES:  4-5
003   T_MATCR3   (3           )  01         TEACHER COURSES:  6-7

CONDITIONING ID:           TMAT0031
DESCRIPTION:               COLLEGE OR IN-SERVICE TRAINING IN SEVEN AREAS
GRADES/ASSESSMENTS:        N04, S04, N08, S08
```

```
GROUP LABEL:              T_MATHTR          LENGTH OF CONTRAST FIELD       :   2
NAEP ID:                  TMATHTR           DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:         CLASS             NUMBER OF SPECIFICATION RECORDS:   3

001   T_MATTR1   (1,M         )  00         TEACHER TRAINING:  0-3
002   T_MATTR2   (2           )  10         TEACHER TRAINING:  4-5
003   T_MATTR3   (3           )  01         TEACHER TRAINING:  6-7
```

APPENDIX D

IRT PARAMETERS FOR MATHEMATICS ITEMS

APPENDIX D

IRT Parameters

This appendix contains 12 tables of IRT (item response theory) parameters for the items that were used in each mathematics scale for the fourth- and eighth-grade Trial State Assessments.

For each of the binary scored items used in scaling (i.e., multiple-choice items and short constructed-response items), the tables provide estimates of the IRT parameters (which correspond to a_j, b_j, and c_j in equation 8.1 in Chapter 8) and their associated standard errors (s.e.) of the estimates. For each of the polytomously scored items (i.e., the extended constructed-response items and the testlets), the tables also show the estimates of the d_{jv} parameters (see equation 8.1) and their associated standard errors.

The tables also show the block in which each item appears for each grade (*Block*) and the position of each item within its block (*Item*).

Note that because the item parameters in this appendix are in the metrics used for the original calibration of the scales, the grade 4 and grade 8 parameters are shown in different metrics. The transformations needed to represent these parameters in terms of the metric of the final reporting scales are given in Chapter 9.

Table D-1
IRT Parameters for Mathematics Items
Numbers and Operations, Grade 4

NAEP ID	*Block*	*Item*	a_j *(s.e.)*	b_j *(s.e.)*	c_j *(s.e.)*	d_{j1} *(s.e.)*	d_{j2} *(s.e.)*	d_{j3} *(s.e.)*	d_{j4} *(s.e.)*
M010131	MH	2	0.404 (0.024)	-2.153 (0.200)	0.199 (0.050)				
M010231	MH	3	0.495 (0.028)	-0.622 (0.125)	0.150 (0.038)				
M010431	MH	5	0.571 (0.041)	0.197 (0.106)	0.206 (0.033)				
M010531	MH	6	0.438 (0.044)	1.720 (0.100)	0.112 (0.023)				
M010631	MH	7	0.334 (0.018)	-1.865 (0.102)	0.000 (0.000)				
M010831	MH	9	0.954 (0.044)	0.148 (0.041)	0.144 (0.018)				
M011131	MH	13	0.860 (0.048)	0.132 (0.059)	0.224 (0.023)				
M017401	MD	1	0.334 (0.024)	-3.176 (0.292)	0.202 (0.056)				
M017701	MD	4	0.956 (0.051)	0.511 (0.040)	0.176 (0.016)				
M017901	MD	6	1.183 (0.055)	0.724 (0.026)	0.101 (0.010)				
M018201	MD	9	0.994 (0.062)	1.277 (0.036)	0.119 (0.010)				
M018401	MD	11	1.290 (0.078)	0.932 (0.031)	0.234 (0.011)				
M018501	MD	12	1.518 (0.187)	2.747 (0.155)	0.255 (0.006)				
M018601	MD	13	0.602 (0.095)	2.735 (0.201)	0.146 (0.013)				
M020001	MF	4	1.032 (0.027)	-0.033 (0.018)	0.000 (0.000)				
M020101	MF	5	0.856 (0.030)	1.516 (0.040)	0.000 (0.000)				
M020501	MF	9	0.656 (0.023)	1.217 (0.040)	0.000 (0.000)				
M021901	ME	1	0.966 (0.046)	-0.070 (0.048)	0.192 (0.021)				
M02200.	ME	2	1.611 (0.077)	1.442 (0.030)	0.188 (0.006)				
M022301	ME	5	0.767 (0.037)	-0.347 (0.066)	0.150 (0.027)				
M022701	ME	9	0.864 (0.078)	1.221 (0.054)	0.306 (0.015)				
M022901	ME	12	1.297 (0.090)	1.452 (0.042)	0.293 (0.009)				
M023001	ME	13	1.515 (0.087)	1.324 (0.033)	0.265 (0.008)				
M039001	MC	1	0.857 (0.040)	-1.629 (0.094)	0.189 (0.045)				
M039201	MC	3	0.752 (0.022)	-0.122 (0.023)	0.000 (0.000)				
M039901	MC	10	0.891 (0.072)	1.729 (0.057)	0.122 (0.009)				
M040201	MC	13	0.961 (0.042)	1.914 (0.053)	0.000 (0.000)				
M040301	MI	1	0.646 (0.020)	0.278 (0.026)	0.000 (0.000)				
M040701	MI	5	1.035 (0.069)	0.839 (0.042)	0.288 (0.014)				
M040901	MI	7	1.294 (0.033)	0.494 (0.016)	0.000 (0.000)				
M041301	ML	1	0.531 (0.025)	-2.545 (0.101)	0.000 (0.000)				
M041401	ML	2	0.577 (0.037)	-0.140 (0.108)	0.181 (0.035)				
M041501	ML	3	1.119 (0.045)	0.029 (0.032)	0.130 (0.015)				
M041701	ML	5	0.420 (0.017)	-0.420 (0.040)	0.000 (0.000)				
M041901	ML	7	0.587 (0.020)	0.350 (0.029)	0.000 (0.000)				
M042401	ML	11	0.868 (0.032)	1.388 (0.038)	0.000 (0.000)				
M042601	MM	1	0.820 (0.039)	-0.404 (0.066)	0.177 (0.028)				
M042901	MM	4	0.512 (0.043)	0.951 (0.088)	0.150 (0.027)				
M043001	MM	5	0.794 (0.046)	-0.566 (0.096)	0.288 (0.036)				
M043301	MM	8	0.760 (0.024)	0.911 (0.029)	0.000 (0.000)				
M043601	MN	1	0.246 (0.023)	-5.469 (0.496)	0.000 (0.000)				
M044001	MN	5	0.861 (0.060)	1.068 (0.045)	0.207 (0.014)				
M044101	MN	6	0.761 (0.063)	1.057 (0.059)	0.266 (0.018)				
M044301	MN	9	1.817 (0.077)	1.911 (0.036)	0.059 (0.004)				
M044401	MN	10	0.358 (0.007)	1.219 (0.024)	0.000 (0.000)	-1.408 (0.079)	0.732 (0.105)	1.410 (0.097)	-0.734 (0.100)
M044501	MG	1	0.613 (0.042)	0.501 (0.078)	0.166 (0.026)				
M044901	MG	5	0.697 (0.040)	-0.417 (0.095)	0.208 (0.035)				
M045001	MG	6	1.466 (0.072)	1.705 (0.036)	0.106 (0.005)				
M045101	MG	7	0.757 (0.027)	1.416 (0.040)	0.000 (0.000)				

Table D-1 (continued)
IRT Parameters for Mathematics Items
Numbers and Operations, Grade 4

NAEP ID	*Block*	*Item*	a_j *(s.e.)*	b_j *(s.e.)*	c_j *(s.e.)*	d_{j1} *(s.e.)*	d_{j2} *(s.e.)*	d_{j3} *(s.e.)*	d_{j4} *(s.e.)*
M045401	MG	10	0.249 (0.005)	1.103 (0.029)	0.000 (0.000)	-1.452 (0.087)	-3.000 (0.186)	2.037 (0.211)	2.415 (0.149)
M046001	MK	1	0.521 (0.019)	-1.041 (0.044)	0.000 (0.000)				
M046301	MK	4	1.155 (0.063)	0.779 (0.031)	0.190 (0.012)				
M046501	MK	6	0.899 (0.087)	1.427 (0.056)	0.306 (0.013)				
M046801	MK	9	1.117 (0.031)	0.774 (0.020)	0.000 (0.000)				
M046901	MK	10	0.938 (0.026)	0.553 (0.021)	0.000 (0.000)				
M047501	MK	16	0.498 (0.030)	-0.713 (0.138)	0.168 (0.042)				
M048101	MO	1	0.553 (0.034)	-0.077 (0.105)	0.164 (0.033)				
M048301	MO	3	0.678 (0.045)	-0.358 (0.117)	0.291 (0.038)				
M048601	MO	6	0.725 (0.040)	0.059 (0.068)	0.168 (0.025)				
M048901	MO	9	0.867 (0.029)	1.327 (0.034)	0.000 (0.000)				
N202831	MH	12	0.664 (0.047)	-0.269 (0.121)	0.283 (0.039)				
N240031	MH	14	1.344 (0.059)	0.460 (0.025)	0.124 (0.011)				
N277903	MF	10	0.561 (0.021)	-1.240 (0.048)	0.000 (0.000)				

Table D-2
IRT Parameters for Mathematics Items
Measurement, Grade 4

NAEP ID	*Block*	*Item*	a_j *(s.e.)*	b_j *(s.e.)*	c_j *(s.e.)*	d_{j1} *(s.e.)*	d_{j2} *(s.e.)*	d_{j3} *(s.e.)*	d_{j4} *(s.e.)*
M010731	MH	8	1.373 (0.085)	0.890 (0.030)	0.279 (0.011)				
M010931	MH	10	1.098 (0.085)	1.316 (0.042)	0.254 (0.011)				
M017501	MD	2	0.575 (0.028)	-1.207 (0.114)	0.146 (0.04[illegible])				
M018101	MD	8	1.254 (0.082)	1.114 (0.033)	0.236 (0.010)				
M020301	MF	7	1.045 (0.032)	1.110 (0.026)	0.000 (0.000)				
M022601	ME	8	0.293 (0.032)	2.074 (0.193)	0.225 (0.024)				
M022801	ME	10	1.511 (0.039)	0.077 (0.014)	0.000 (0.000)				
M022802	ME	11	1.380 (0.035)	-0.161 (0.015)	0.000 (0.000)				
M023401	ME	17	1.636 (0.095)	1.793 (0.046)	0.138 (0.006)				
M039101	MC	2	0.481 (0.026)	-1.269 (0.144)	0.155 (0.044)				
M039301	MC	4	0.911 (0.027)	0.854 (0.024)	0.000 (0.000)				
M039401	MC	5	0.676 (0.042)	0.006 (0.083)	0.196 (0.029)				
M039501	MC	6	0.904 (0.038)	-0.327 (0.047)	0.113 (0.021)				
M039601	MC	7	0.678 (0.041)	0.648 (0.054)	0.116 (0.019)				
M040461	MI	2	0.474 (0.008)	-0.991 (0.021)	0.000 (0.000)	-0.346 (0.062)	-1.963 (0.089)	2.309 (0.079)	**** (****)
M040801	MI	6	0.865 (0.059)	1.250 (0.043)	0.152 (0.012)				
M041001	MI	8	1.301 (0.076)	1.536 (0.035)	0.108 (0.006)				
M041601	ML	4	0.866 (0.060)	1.158 (0.043)	0.173 (0.013)				
M042701	MM	2	0.920 (0.038)	-1.522 (0.071)	0.137 (0.037)				
M042801	MM	3	0.769 (0.035)	-1.731 (0.097)	0.257 (0.045)				
M043701	MN	2	1.185 (0.053)	0.486 (0.027)	0.124 (0.011)				
M044601	MG	2	0.557 (0.020)	0.466 (0.031)	0.000 (0.000)				
M047101	MK	12	1.414 (0.086)	1.934 (0.051)	0.170 (0.006)				
M047201	MK	13	0.963 (0.060)	0.989 (0.037)	0.161 (0.012)				
M048201	MO	2	0.490 (0.028)	-0.921 (0.142)	0.174 (0.043)				
M048401	MO	4	0.908 (0.049)	0.647 (0.038)	0.147 (0.014)				
M048501	MO	5	0.530 (0.033)	-0.040 (0.103)	0.137 (0.033)				
M048701	MO	7	0.931 (0.030)	1.275 (0.031)	0.000 (0.000)				
M061906	MJ	6	0.942 (0.051)	2.516 (0.090)	0.000 (0.000)				

Table D-3
IRT Parameters for Mathematics Items
Geometry, Grade 4

NAEP ID	*Block*	*Item*	a_j *(s.e.)*	b_j *(s.e.)*	c_j *(s.e.)*	d_{j1} *(s.e.)*	d_{j2} *(s.e.)*	d_{j3} *(s.e.)*	d_{j4} *(s.e.)*
M011231	MH	15	1.370 (0.097)	1.999 (0.062)	0.241 (0.007)				
M017601	MD	3	0.307 (0.022)	-0.625 (0.177)	0.247 (0.033)				
M018001	MD	7	0.885 (0.068)	1.438 (0.049)	0.185 (0.012)				
M019801	MF	2	0.353 (0.025)	3.718 (0.238)	0.000 (0.000)				
M019901	MF	3	0.602 (0.020)	-0.742 (0.033)	0.000 (0.000)				
M020701	MF	11	0.569 (0.023)	1.353 (0.052)	0.000 (0.000)				
M022201	ME	4	0.663 (0.021)	0.462 (0.027)	0.000 (0.000)				
M022501	ME	7	0.935 (0.030)	1.288 (0.031)	0.000 (0.000)				
M023101	ME	14	1.257 (0.086)	1.448 (0.037)	0.136 (0.008)				
M039801	MC	9	0.647 (0.086)	1.935 (0.105)	0.329 (0.017)				
M041201	MI	10	0.307 (0.008)	1.959 (0.032)	0.000 (0.000)	2.515 (0.063)	1.134 (0.066)	-0.538 (0.097)	-3.111 (0.252)
M041801	ML	6	0.900 (0.067)	1.740 (0.054)	0.103 (0.009)				
M043401	MM	9	2.314 (0.062)	0.351 (0.011)	0.000 (0.000)				
M043402	MM	10	2.383 (0.066)	0.528 (0.011)	0.000 (0.000)				
M043403	MM	11	1.213 (0.053)	2.020 (0.050)	0.000 (0.000)				
M043901	MN	4	0.537 (0.039)	0.102 (0.121)	0.201 (0.036)				
M044801	MG	4	0.559 (0.043)	-0.318 (0.154)	0.270 (0.045)				
M046101	MK	2	0.603 (0.035)	-1.027 (0.139)	0.223 (0.048)				
M046201	MK	3	1.038 (0.060)	0.838 (0.035)	0.183 (0.013)				
M046401	MK	5	0.733 (0.048)	0.446 (0.064)	0.198 (0.023)				
M046701	MK	8	2.497 (0.058)	1.886 (0.103)	0.172 (0.022)				
M047401	MK	15	0.890 (0.043)	-1.696 (0.095)	0.198 (0.048)				
M061901	MJ	1	0.532 (0.019)	-0.157 (0.030)	0.000 (0.000)				
M061902	MJ	2	1.514 (0.038)	-0.015 (0.014)	0.000 (0.000)				
M061903	MJ	3	1.714 (0.044)	-0.403 (0.014)	0.000 (0.000)				
M061904	MJ	4	0.833 (0.032)	1.851 (0.052)	0.000 (0.000)				
N214331	MH	1	0.817 (0.041)	-2.034 (0.112)	0.210 (0.053)				

Table D-4
IRT Parameters for Mathematics Items
Data Analysis, Statistics, and Probability, Grade 4

NAEP ID	*Block*	*Item*	a_j (*s.e.*)	b_j (*s.e.*)	c_j (*s.e.*)	d_{j1} (*s.e.*)	d_{j2} (*s.e.*)	d_{j3} (*s.e.*)	d_{j4} (*s.e.*)
M017801	MD	5	0.912 (0.057)	0.800 (0.044)	0.213 (0.015)				
M020201	MF	6	0.865 (0.024)	-0.091 (0.020)	0.000 (0.000)				
M023301	ME	16	2.204 (0.090)	1.400 (0.026)	0.177 (0.006)				
M040001	MC	11	1.090 (0.028)	0.310 (0.018)	0.000 (0.000)				
M040101	MC	12	0.788 (0.069)	0.884 (0.073)	0.392 (0.020)				
M040601	MI	4	0.950 (0.045)	0.048 (0.045)	0.166 (0.020)				
M041101	MI	9	2.073 (0.090)	1.846 (0.036)	0.240 (0.006)				
M042001	ML	8	0.723 (0.041)	-0.159 (0.079)	0.188 (0.030)				
M042002	ML	9	0.583 (0.020)	-0.137 (0.029)	0.000 (0.000)				
M042003	ML	10	1.133 (0.031)	0.077 (0.018)	0.000 (0.000)				
M043101	MM	6	1.071 (0.066)	0.739 (0.040)	0.266 (0.014)				
M043201	MM	7	0.692 (0.021)	-0.615 (0.028)	0.000 (0.000)				
M044701	MG	3	0.559 (0.081)	1.846 (0.118)	0.367 (0.022)				
M045301	MG	9	1.067 (0.032)	1.027 (0.025)	0.000 (0.000)				
M046601	MK	7	1.188 (0.032)	0.667 (0.018)	0.000 (0.000)				
M047001	MK	11	1.834 (0.081)	1.864 (0.037)	0.161 (0.005)				
M047301	MK	14	1.276 (0.033)	-0.168 (0.016)	0.000 (0.000)				
M049001	MO	10	0.347 (0.010)	1.754 (0.037)	0.000 (0.000)	0.517 (0.056)	0.928 (0.071)	-1.016 (0.111)	-0.429 (0.164)
M061905	MJ	5	0.878 (0.027)	0.987 (0.028)	0.000 (0.000)				
N250231	MH	11	0.813 (0.044)	0.288 (0.053)	0.160 (0.021)				

Table D-5
IRT Parameters for Mathematics Items
Algebra and Functions, Grade 4

NAEP ID	*Block*	*Item*	a_j *(s.e.)*	b_j *(s.e.)*	c_j *(s.e.)*	d_{j1} *(s.e.)*	d_{j2} *(s.e.)*	d_{j3} *(s.e.)*	d_{j4} *(s.e.)*
M010331	MH	4	0.633 (0.044)	0.133 (0.100)	0.224 (0.033)				
M018301	MD	10	1.201 (0.062)	0.998 (0.029)	0.129 (0.009)				
M018701	MD	14	1.033 (0.097)	2.644 (0.110)	0.173 (0.006)				
M019701	MF	1	0.686 (0.021)	-0.200 (0.024)	0.000 (0.000)				
M020401	MF	8	0.607 (0.020)	0.553 (0.031)	0.000 (0.000)				
M022101	ME	3	0.707 (0.037)	-2.011 (0.127)	0.202 (0.055)				
M022401	ME	6	0.494 (0.046)	0.775 (0.127)	0.227 (0.034)				
M023201	ME	15	1.057 (0.084)	1.503 (0.048)	0.197 (0.010)				
M039701	MC	8	0.665 (0.054)	1.515 (0.061)	0.142 (0.016)				
M040501	MI	3	0.694 (0.041)	-0.252 (0.093)	0.205 (0.034)				
M042501	ML	12	0.756 (0.028)	1.207 (0.041)	0.000 (0.000)				
M043501	MM	12	0.312 (0.009)	1.112 (0.030)	0.000 (0.000)	0.988 (0.061)	-0.509 (0.080)	0.176 (0.099)	-0.654 (0.118)
M043801	MN	3	0.741 (0.022)	-0.021 (0.023)	0.000 (0.000)				
M044261	MN		0.544 (0.015)	0.734 (0.025)	0.000 (0.000)	-0.003 (0.038)	0.003 (0.046)	**** (****)	**** (****)
M045201	MG	8	0.880 (0.094)	2.183 (0.084)	0.188 (0.009)				
M048801	MO	8	0.881 (0.025)	0.526 (0.022)	0.000 (0.000)				

BEST COPY AVAILABLE

Table D-6
IRT Parameters for Mathematics Items
Estimation, Grade 4

NAEP ID	*Block*	*Item*	a_j *(s.e.)*	b_j *(s.e.)*	c_j *(s.e.)*	d_{j1} *(s.e.)*	d_{j2} *(s.e.)*	d_{j3} *(s.e.)*	d_{j4} *(s.e.)*
M031101	MP	11	0.863 (0.033)	-0.861 (0.075)	0.532 (0.022)				
M031201	MP	12	1.126 (0.063)	1.366 (0.032)	0.597 (0.005)				
M031301	MP	13	0.831 (0.024)	-1.293 (0.067)	0.327 (0.029)				
M031401	MP	14	0.415 (0.023)	0.990 (0.076)	0.244 (0.019)				
M031402	MP	15	0.625 (0.020)	1.019 (0.025)	0.072 (0.008)				
M031501	MP	16	0.992 (0.038)	0.872 (0.024)	0.362 (0.007)				
M031601	MP	17	0.506 (0.014)	-0.774 (0.068)	0.214 (0.021)				
M031701	MP	18	0.690 (0.021)	-1.071 (0.073)	0.358 (0.025)				
M031801	MP	19	1.320 (0.045)	0.859 (0.017)	0.383 (0.006)				
M031901	MP	20	0.103 (0.010)	2.758 (0.328)	0.363 (0.011)				
M032001	MP	1	0.716 (0.025)	0.824 (0.027)	0.162 (0.010)				
M032101	MP	2	1.210 (0.094)	2.521 (0.091)	0.491 (0.003)				
M032201	MP	3	0.439 (0.023)	0.658 (0.081)	0.286 (0.019)				
M032301	MP	4	0.541 (0.026)	0.955 (0.046)	0.188 (0.015)				
M032401	MP	5	0.488 (0.030)	1.555 (0.047)	0.183 (0.014)				
M032501	MP	6	0.683 (0.031)	1.428 (0.029)	0.173 (0.008)				
M032601	MP	7	0.735 (0.044)	1.509 (0.036)	0.356 (0.008)				
M032701	MP	8	0.318 (0.012)	-0.142 (0.085)	0.341 (0.015)				
M032901	MP	10	1.288 (0.046)	2.055 (0.033)	0.210 (0.003)				

Table D-7
IRT Parameters for Mathematics Items
Numbers and Operations, Grade 8

NAEP ID	*Block*	*Item*	a_j *(s.e.)*	b_j *(s.e.)*	c_j *(s.e.)*	d_{j1} *(s.e.)*	d_{j2} *(s.e.)*	d_{j3} *(s.e.)*	d_{j4} *(s.e.)*
M011131	MH	13	0.626 (0.034)	-1.467 (0.136)	0.217 (0.049)				
M012431	MH	3	0.882 (0.037)	-0.329 (0.048)	0.109 (0.021)				
M012531	MH	4	0.729 (0.040)	0.647 (0.044)	0.092 (0.016)				
M012931	MH	8	1.106 (0.074)	1.023 (0.035)	0.227 (0.012)				
M013431	MH	15	1.152 (0.057)	0.165 (0.037)	0.179 (0.017)				
M013531	MH	16	1.048 (0.090)	1.561 (0.051)	0.141 (0.010)				
M013631	MH	17	1.326 (0.064)	0.871 (0.024)	0.053 (0.007)				
M017401	MD	1	0.235 (0.027)	-5.827 (0.707)	0.247 (0.063)				
M017701	MD	4	0.910 (0.036)	-1.075 (0.059)	0.130 (0.027)				
M017901	MD	6	1.179 (0.050)	-0.881 (0.048)	0.173 (0.025)				
M018201	MD	9	0.563 (0.027)	-0.807 (0.097)	0.136 (0.032)				
M018401	MD	11	1.011 (0.047)	-0.994 (0.067)	0.220 (0.032)				
M018501	MD	12	1.975 (0.101)	0.564 (0.021)	0.270 (0.010)				
M018601	MD	13	0.682 (0.050)	1.027 (0.053)	0.143 (0.017)				
M020001	MF	4	0.603 (0.020)	-0.483 (0.031)	0.000 (0.000)				
M020101	MF	5	1.293 (0.034)	-0.382 (0.017)	0.000 (0.000)				
M020501	MF	9	0.755 (0.022)	-0.309 (0.024)	0.000 (0.000)				
M021901	ME	1	0.907 (0.039)	-1.403 (0.074)	0.167 (0.035)				
M022001	ME	2	1.180 (0.052)	-0.649 (0.046)	0.212 (0.023)				
M022301	ME	5	0.557 (0.032)	-2.609 (0.184)	0.235 (0.060)				
M022701	ME	9	0.991 (0.048)	-0.635 (0.062)	0.256 (0.027)				
M022901	ME	12	1.096 (0.056)	-0.282 (0.053)	0.297 (0.022)				
M023001	ME	13	1.233 (0.062)	-0.146 (0.043)	0.297 (0.019)				
M023801	ME	21	1.599 (0.069)	0.327 (0.022)	0.103 (0.011)				
M044501	MG	1	0.553 (0.033)	-1.253 (0.159)	0.232 (0.050)				
M044901	MG	5	0.817 (0.043)	-2.123 (0.129)	0.250 (0.058)				
M045001	MG	6	1.302 (0.053)	-0.187 (0.032)	0.168 (0.016)				
M045101	MG	7	0.657 (0.020)	-0.462 (0.028)	0.000 (0.000)				
M046001	MK	1	0.330 (0.020)	-3.203 (0.184)	0.000 (0.000)				
M046301	MK	4	0.999 (0.048)	-0.914 (0.070)	0.250 (0.032)				
M046501	MK	6	0.888 (0.058)	-0.038 (0.074)	0.374 (0.025)				
M046801	MK	9	0.619 (0.021)	-1.252 (0.043)	0.000 (0.000)				
M046901	MK	10	0.938 (0.029)	-1.406 (0.034)	0.000 (0.000)				
M048101	MO	1	0.377 (0.027)	-3.423 (0.292)	0.231 (0.060)				
M048301	MO	3	0.573 (0.031)	-2.235 (0.164)	0.223 (0.056)				
M048601	MO	6	0.924 (0.044)	-1.824 (0.094)	0.205 (0.046)				
M048901	MO	9	0.728 (0.022)	0.169 (0.023)	0.000 (0.000)				
M049101	MO	10	1.187 (0.061)	1.238 (0.028)	0.043 (0.006)				
M049901	MC	1	0.718 (0.038)	-0.562 (0.087)	0.175 (0.034)				
M050001	MC	2	1.220 (0.046)	-0.310 (0.032)	0.109 (0.016)				
M050101	MC	3	0.978 (0.044)	0.590 (0.029)	0.071 (0.011)				
M050301	MC	5	1.373 (0.085)	1.193 (0.029)	0.186 (0.009)				
M051101	MC	13	0.371 (0.010)	1.608 (0.027)	0.000 (0.000)	1.438 (0.055)	1.135 (0.059)	-0.288 (0.078)	-2.285 (0.174)
M051201	MM	1	0.509 (0.021)	-1.904 (0.072)	0.000 (0.000)				
M051501	MM	4	0.938 (0.096)	1.785 (0.069)	0.191 (0.010)				
M051601	MM	5	0.785 (0.024)	-0.894 (0.030)	0.000 (0.000)				
M051901	MM	8	1.504 (0.061)	0.766 (0.018)	0.054 (0.006)				
M052401	MI	2	0.984 (0.029)	0.781 (0.022)	0.000 (0.000)				
M052901	MI	7	0.758 (0.023)	-0.009 (0.023)	0.000 (0.000)				

Table D-7 (continued)
IRT Parameters for Mathematics Items
Numbers and Operations, Grade 8

NAEP ID	*Block*	*Item*	a_j *(s.e.)*	b_j *(s.e.)*	c_j *(s.e.)*	d_{j1} *(s.e.)*	d_{j2} *(s.e.)*	d_{j3} *(s.e.)*	d_{j4} *(s.e.)*
M053001	MI	8	0.814 (0.030)	1.336 (0.037)	0.000 (0.000)				
M053601	ML	2	0.655 (0.037)	-0.140 (0.080)	0.146 (0.029)				
M053701	ML	3	0.960 (0.081)	0.844 (0.054)	0.400 (0.016)				
M053901	ML	5	1.139 (0.062)	0.779 (0.030)	0.157 (0.012)				
M054701	MN	1	0.543 (0.038)	-0.911 (0.174)	0.301 (0.049)				
M054801	MN	2	0.602 (0.020)	0.088 (0.027)	0.000 (0.000)				
M055201	MN	6	1.102 (0.029)	-0.022 (0.017)	0.000 (0.000)				
M055501	MN	9	0.415 (0.013)	1.752 (0.034)	0.000 (0.000)	1.191 (0.052)	0.736 (0.062)	-1.834 (0.135)	-0.093 (0.204)
N202831	MH	12	0.647 (0.036)	-1.996 (0.157)	0.246 (0.058)				

Table D-8
IRT Parameters for Mathematics Items
Measurement, Grade 8

NAEP ID	*Block*	*Item*	a_j *(s.e.)*	b_j *(s.e.)*	c_j *(s.e.)*	d_{j1} *(s.e.)*	d_{j2} *(s.e.)*	d_{j3} *(s.e.)*	d_{j4} *(s.e.)*
M012331	MH	2	0.716 (0.033)	-1.523 (0.103)	0.164 (0.045)				
M013331	MH	14	0.868 (0.040)	-1.337 (0.081)	0.153 (0.039)				
M017501	MD	2	0.375 (0.027)	-2.960 (0.269)	0.257 (0.056)				
M018101	MD	8	0.599 (0.032)	-0.555 (0.099)	0.170 (0.034)				
M019101	MD	18	1.173 (0.083)	1.874 (0.054)	0.154 (0.007)				
M019201	MD	19	1.288 (0.081)	1.850 (0.049)	0.141 (0.006)				
M020301	MF	7	1.082 (0.029)	-0.357 (0.019)	0.000 (0.000)				
M022601	ME	8	1.157 (0.079)	0.717 (0.042)	0.357 (0.014)				
M022801	ME	10	1.623 (0.044)	-0.684 (0.015)	0.000 (0.000)				
M022802	ME	11	1.422 (0.042)	-1.061 (0.020)	0.000 (0.000)				
M023401	ME	17	1.291 (0.081)	0.318 (0.043)	0.385 (0.016)				
M023701	ME	20	0.653 (0.024)	0.708 (0.033)	0.000 (0.000)				
M044601	MG	2	0.681 (0.021)	-0.584 (0.028)	0.000 (0.000)				
M047101	MK	12	1.339 (0.059)	0.342 (0.027)	0.165 (0.012)				
M047201	MK	13	0.865 (0.041)	-0.696 (0.071)	0.201 (0.031)				
M047901	MK	18	0.840 (0.033)	1.861 (0.052)	0.000 (0.000)				
M048201	MO	2	0.567 (0.041)	-0.531 (0.156)	0.295 (0.045)				
M048401	MO	4	0.967 (0.045)	-0.932 (0.068)	0.202 (0.033)				
M048501	MO	5	0.810 (0.045)	-0.898 (0.101)	0.272 (0.041)				
M048701	MO	7	0.572 (0.020)	-0.619 (0.033)	0.000 (0.000)				
M049201	MO	11	0.718 (0.060)	1.015 (0.063)	0.249 (0.019)				
M049501	MO	14	0.792 (0.062)	1.367 (0.051)	0.160 (0.014)				
M050501	MC	7	0.717 (0.049)	0.246 (0.079)	0.247 (0.027)				
M050901	MC	11	0.876 (0.030)	1.386 (0.036)	0.000 (0.000)				
M051301	MM	2	0.344 (0.022)	-3.354 (0.204)	0.000 (0.000)				
M052201	MM	11	0.571 (0.014)	1.229 (0.021)	0.000 (0.000)	1.021 (0.033)	-0.427 (0.050)	-1.061 (0.092)	0.466 (0.103)
M052301	MI	1	1.382 (0.080)	1.102 (0.028)	0.179 (0.009)				
M054001	ML	6	0.959 (0.040)	2.053 (0.055)	0.000 (0.000)				
M055101	MN	5	0.665 (0.071)	1.952 (0.087)	0.163 (0.014)				
M055401	MN	8	0.804 (0.056)	1.254 (0.047)	0.130 (0.013)				
M061907	MJ	5	0.940 (0.028)	0.962 (0.026)	0.000 (0.000)				
M061908	MJ	6	0.752 (0.037)	2.448 (0.088)	0.000 (0.000)				

Table D-9
IRT Parameters for Mathematics Items
Geometry, Grade 8

NAEP ID	*Block*	*Item*	a_j (s.e.)	b_j (s.e.)	c_j (s.e.)	d_{j1} (s.e.)	d_{j2} (s.e.)	d_{j3} (s.e.)	d_{j4} (s.e.)
M012731	MH	6	0.901 (0.077)	1.490 (0.054)	0.233 (0.012)				
M012831	MH	7	1.238 (0.062)	0.737 (0.027)	0.142 (0.010)				
M017601	MD	3	0.537 (0.032)	-1.416 (0.161)	0.270 (0.049)				
M018001	MD	7	0.860 (0.047)	0.086 (0.057)	0.216 (0.022)				
M019001	MD	17	0.699 (0.044)	0.747 (0.052)	0.117 (0.019)				
M019601	MD	21	0.635 (0.056)	1.573 (0.069)	0.119 (0.016)				
M019801	MF	2	0.857 (0.025)	-0.642 (0.025)	0.000 (0.000)				
M019901	MF	3	0.608 (0.023)	-1.693 (0.058)	0.000 (0.000)				
M020901	MF	11	0.606 (0.022)	0.958 (0.037)	0.000 (0.000)				
M021001	MF	12	0.829 (0.024)	0.214 (0.021)	0.000 (0.000)				
M021301	MF	15	1.373 (0.036)	0.043 (0.015)	0.000 (0.000)				
M021302	MF	16	1.289 (0.035)	-0.190 (0.017)	0.000 (0.000)				
M022201	ME	4	0.687 (0.022)	-0.851 (0.032)	0.000 (0.000)				
M022501	ME	7	0.793 (0.023)	-0.387 (0.023)	0.000 (0.000)				
M023101	ME	14	1.231 (0.053)	0.048 (0.031)	0.151 (0.015)				
M044801	MG	4	0.729 (0.039)	-1.641 (0.134)	0.259 (0.055)				
M045601	MG	10	0.278 (0.016)	0.213 (0.056)	0.000 (0.000)				
M045901	MG	13	0.492 (0.014)	1.765 (0.030)	0.000 (0.000)	0.651 (0.043)	0.718 (0.057)	-0.629 (0.091)	-0.740 (0.148)
M046101	MK	2	0.634 (0.036)	-2.466 (0.163)	0.248 (0.060)				
M046201	MK	3	1.094 (0.051)	-0.419 (0.048)	0.212 (0.023)				
M046401	MK	5	0.755 (0.044)	-0.823 (0.109)	0.276 (0.041)				
M046701	MK	8	1.405 (0.073)	0.282 (0.032)	0.295 (0.014)				
M048001	MK	19	1.250 (0.081)	1.500 (0.036)	0.092 (0.007)				
M049301	MO	12	0.778 (0.058)	1.600 (0.055)	0.099 (0.011)				
M049701	MO	16	1.030 (0.061)	1.018 (0.034)	0.142 (0.011)				
M051001	MC	12	0.590 (0.022)	1.078 (0.041)	0.000 (0.000)				
M051801	MM	7	1.422 (0.090)	1.748 (0.048)	0.246 (0.007)				
M052001	MM	9	0.864 (0.065)	1.403 (0.048)	0.151 (0.012)				
M052601	MI	4	0.661 (0.040)	0.446 (0.061)	0.126 (0.022)				
M054101	ML	7	0.776 (0.023)	-0.137 (0.023)	0.000 (0.000)				
M054201	ML	8	1.009 (0.089)	1.613 (0.054)	0.218 (0.010)				
M055301	MN	7	0.744 (0.245)	4.126 (0.743)	0.241 (0.008)				
M061901	MJ	1	0.597 (0.022)	-1.315 (0.047)	0.000 (0.000)				
M061902	MJ	4	1.480 (0.043)	-0.963 (0.019)	0.000 (0.000)				
M061903	MJ	2	1.318 (0.042)	-1.399 (0.027)	0.000 (0.000)				
M061904	MJ	3	0.990 (0.027)	0.256 (0.019)	0.000 (0.000)				

Table D-10
IRT Parameters for Mathematics Items
Data Analysis, Statistics, and Probability, Grade 8

NAEP ID	*Block*	*Item*	a_j *(s.e.)*	b_j *(s.e.)*	c_j *(s.e.)*	d_{j1} *(s.e.)*	d_{j2} *(s.e.)*	d_{j3} *(s.e.)*	d_{j4} *(s.e.)*
M012631	MH	5	1.189 (0.062)	0.597 (0.031)	0.201 (0.013)				
M013031	MH	9	0.991 (0.036)	1.513 (0.036)	0.000 (0.000)				
M013131	MH	10	0.851 (0.032)	1.563 (0.042)	0.000 (0.000)				
M017801	MD	5	1.390 (0.071)	-0.226 (0.043)	0.356 (0.019)				
M018901	MD	16	0.911 (0.111)	2.323 (0.127)	0.157 (0.008)				
M020201	MF	6	0.656 (0.026)	-2.072 (0.066)	0.000 (0.000)				
M020801	MF	10	0.924 (0.038)	1.789 (0.049)	0.000 (0.000)				
M021101	MF	13	0.914 (0.025)	0.104 (0.020)	0.000 (0.000)				
M023301	ME	16	1.787 (0.084)	-0.425 (0.032)	0.273 (0.018)				
M023501	ME	18	1.623 (0.083)	0.871 (0.021)	0.132 (0.008)				
M023601	ME	19	0.926 (0.043)	-0.347 (0.054)	0.133 (0.024)				
M044701	MG	3	0.890 (0.070)	0.615 (0.062)	0.387 (0.019)				
M045301	MG	9	0.956 (0.025)	-0.260 (0.020)	0.000 (0.000)				
M045861	MG	12	0.446 (0.010)	-0.765 (0.018)	0.000 (0.000)	1.630 (0.075)	-0.645 (0.057)	-0.659 (0.059)	-0.326 (0.050)
M046601	MK	7	1.156 (0.032)	-0.927 (0.022)	0.000 (0.000)				
M047001	MK	11	0.998 (0.045)	0.223 (0.037)	0.127 (0.015)				
M047301	MK	14	1.017 (0.036)	-1.755 (0.040)	0.000 (0.000)				
M047801	MK	17	1.524 (0.079)	0.892 (0.022)	0.148 (0.009)				
M049801	MO	17	0.796 (0.043)	2.246 (0.086)	0.000 (0.000)				
M050261	MC	4	0.536 (0.011)	-2.003 (0.026)	0.000 (0.000)	1.225 (0.128)	-1.552 (0.098)	-0.081 (0.092)	0.408 (0.050)
M050401	MC	6	0.626 (0.033)	-0.328 (0.089)	0.131 (0.032)				
M051401	MM	3	0.940 (0.038)	-0.963 (0.058)	0.121 (0.029)				
M052701	MI	5	0.988 (0.048)	0.503 (0.034)	0.128 (0.014)				
M052801	MI	6	1.178 (0.246)	3.002 (0.320)	0.203 (0.006)				
M053101	MI	9	0.603 (0.014)	1.347 (0.019)	0.000 (0.000)	0.420 (0.033)	-1.295 (0.081)	1.303 (0.089)	-0.429 (0.078)
M053801	ML	4	0.435 (0.112)	3.512 (0.480)	0.233 (0.019)				
M054901	MN	3	0.895 (0.107)	2.126 (0.103)	0.146 (0.009)				
M061905	MJ	7	0.532 (0.020)	-0.462 (0.035)	0.000 (0.000)				

Table D-11
IRT Parameters for Mathematics Items
Algebra and Functions, Grade 8

NAEP ID	*Block*	*Item*	a_j *(s.e.)*	b_j *(s.e.)*	c_j *(s.e.)*	d_{j1} *(s.e.)*	d_{j2} *(s.e.)*	d_{j3} *(s.e.)*	d_{j4} *(s.e.)*
M012231	MH	1	0.579 (0.036)	-3.092 (0.183)	0.166 (0.053)				
M013231	MH	11	1.007 (0.087)	1.839 (0.060)	0.130 (0.008)				
M013731	MH	18	1.148 (0.088)	1.299 (0.041)	0.135 (0.011)				
M018301	MD	10	0.939 (0.041)	-0.504 (0.054)	0.158 (0.025)				
M018701	MD	14	1.705 (0.081)	0.320 (0.025)	0.251 (0.012)				
M018801	MD	15	1.246 (0.089)	1.011 (0.036)	0.315 (0.012)				
M019301	MD	20	1.653 (0.084)	1.238 (0.028)	0.209 (0.008)				
M019701	MF	1	0.439 (0.020)	-1.933 (0.083)	0.000 (0.000)				
M020401	MF	8	0.634 (0.020)	-0.010 (0.026)	0.000 (0.000)				
M021201	MF	14	0.987 (0.028)	0.573 (0.021)	0.000 (0.000)				
M022101	ME	3	0.721 (0.044)	-2.819 (0.158)	0.231 (0.059)				
M022401	ME	6	1.164 (0.057)	-0.770 (0.060)	0.277 (0.029)				
M023201	ME	15	1.068 (0.050)	-0.396 (0.051)	0.203 (0.024)				
M045201	MG	8	0.787 (0.052)	0.582 (0.056)	0.215 (0.020)				
M045701	MG	11	1.232 (0.031)	0.096 (0.016)	0.000 (0.000)				
M047601	MK	15	1.354 (0.061)	0.564 (0.024)	0.123 (0.010)				
M047701	MK	16	1.555 (0.090)	1.651 (0.042)	0.260 (0.007)				
M048801	MO	8	1.140 (0.032)	-0.903 (0.022)	0.000 (0.000)				
M049401	MO	13	0.956 (0.055)	1.194 (0.034)	0.076 (0.009)				
M049601	MO	15	0.754 (0.035)	-0.412 (0.066)	0.120 (0.027)				
M050601	MC	8	1.608 (0.074)	0.723 (0.020)	0.125 (0.008)				
M050701	MC	9	1.122 (0.048)	0.046 (0.034)	0.131 (0.015)				
M050801	MC	10	0.736 (0.022)	0.025 (0.023)	0.000 (0.000)				
M051701	MM	6	1.375 (0.072)	0.710 (0.026)	0.194 (0.011)				
M052101	MM	10	0.895 (0.025)	0.376 (0.021)	0.000 (0.000)				
M052501	MI	3	1.419 (0.080)	1.297 (0.030)	0.173 (0.008)				
M053501	ML	1	1.422 (0.056)	-0.432 (0.032)	0.147 (0.018)				
M054301	ML	9	0.413 (0.009)	1.645 (0.032)	0.000 (0.000)	-0.955 (0.061)	0.081 (0.097)	-1.226 (0.174)	2.100 (0.178)
M055001	MN	4	0.634 (0.058)	1.982 (0.084)	0.079 (0.012)				

Table D-12
IRT Parameters for Mathematics Items
Estimation, Grade 8

NAEP ID	*Block*	*Item*	a_j *(s.e.)*	b_j *(s.e.)*	c_j *(s.e.)*	d_{j1} *(s.e.)*	d_{j2} *(s.e.)*	d_{j3} *(s.e.)*	d_{j4} *(s.e.)*
M032001	MP	1	0.602 (0.020)	-0.235 (0.060)	0.200 (0.020)				
M032201	MP	3	0.522 (0.016)	-1.624 (0.087)	0.354 (0.024)				
M032301	MP	4	0.727 (0.024)	-0.442 (0.058)	0.305 (0.020)				
M032401	MP	5	0.466 (0.023)	0.358 (0.085)	0.355 (0.019)				
M032501	MP	6	0.263 (0.010)	-0.419 (0.088)	0.206 (0.015)				
M032601	MP	7	0.788 (0.026)	-0.157 (0.045)	0.327 (0.015)				
M032701	MP	8	0.598 (0.018)	-1.519 (0.083)	0.371 (0.026)				
M032801	MP	9	1.714 (0.048)	0.885 (0.013)	0.324 (0.005)				
M032901	MP	10	0.855 (0.024)	0.119 (0.030)	0.232 (0.011)				
M033001	MP	11	0.759 (0.029)	1.148 (0.024)	0.170 (0.008)				
M033101	MP	12	0.407 (0.012)	-0.248 (0.064)	0.115 (0.017)				
M033201	MP	13	0.778 (0.038)	0.281 (0.056)	0.556 (0.012)				
M033301	MP	14	1.013 (0.022)	0.520 (0.014)	0.071 (0.006)				
M033401	MP	15	0.678 (0.049)	1.574 (0.047)	0.441 (0.009)				
M033501	MP	16	0.474 (0.019)	0.006 (0.077)	0.289 (0.019)				
M033601	MP	17	0.544 (0.036)	2.086 (0.055)	0.153 (0.010)				
M033701	MP	18	1.074 (0.037)	1.238 (0.018)	0.193 (0.005)				
M033801	MP	19	1.217 (0.029)	0.166 (0.018)	0.247 (0.008)				
M033901	MP	20	0.712 (0.039)	1.395 (0.035)	0.322 (0.009)				
M034001	MP	21	0.504 (0.045)	2.500 (0.092)	0.204 (0.011)				
M034101	MP	22	1.153 (0.041)	1.504 (0.019)	0.135 (0.004)				

APPENDIX E

TRIAL STATE ASSESSMENT REPORTING SUBGROUPS

COMPOSITE AND DERIVED COMMON BACKGROUND VARIABLES

COMPOSITE AND DERIVED REPORTING VARIABLES

REPORTING SUBGROUPS FOR THE 1992 TRIAL STATE ASSESSMENT

Results for the 1990 Trial State Assessment were reported for student subgroups defined by gender, race/ethnicity, type of community, parents' level of education, and geographical region. The following explains how each of these subgroups was derived.

DSEX (Gender)

The variable SEX is the gender of the student being assessed, as taken from school records. For a few students, data for this variable was missing and was imputed by ETS after the assessment. The resulting variable DSEX contains a value for every student and is used for gender comparisons among students.

DRACE (Race/ethnicity)

The variable DRACE is an imputed definition of race/ethnicity, derived from up to three sources of information. This variable is used for race/ethnicity subgroup comparisons. Two items from the student demographics questionnaire were used in the determination of derived race/ethnicity:

Demographic Item Number 2:

2. If you are Hispanic, what is your Hispanic background?

○ I am not Hispanic.
○ Mexican, Mexican American, or Chicano
○ Puerto Rican
○ Cuban
○ Other Spanish or Hispanic background

Students who responded to item number 2 by filling in the second, third, fourth, or fifth oval were considered Hispanic. For students who filled in the first oval, did not respond to the item, or provided information that was illegible or could not be classified, responses to item number 1 were examined in an effort to determine race/ethnicity. Item number 1 read as follows:

Demographic Item Number 1:

1. Which best describes you?

○ White (not Hispanic)

○ Black (not Hispanic)

○ Hispanic ("Hispanic" means someone who is Mexican, Mexican American, Chicano, Puerto Rican, Cuban, or from some other Spanish or Hispanic background.)

○ Asian or Pacific Islander ("Asian or Pacific Islander" means someone who is Chinese, Japanese, Korean, Filipino, Vietnamese, or from some other Asian or Pacific Island background.)

○ American Indian or Alaskan Native ("American Indian or Alaskan Native" means someone who is from one of the American Indian tribes, or one of the original people of Alaska.)

○ Other (What?) ________

Students' race/ethnicity was then assigned to correspond with their selection. For students who filled in the sixth oval ("Other"), provided illegible information or information that could not be classified, or did not respond at all, race/ethnicity as provided from school records was used.

Derived race/ethnicity could not be determined for students who did not respond to background items 1 or 2 and for whom race/ethnicity was not provided by the school.

TOC (Type of community)

NAEP assigned each participating school to one of four type of categories designed to provide information about the communities in which the schools are located.

The type of community categories consist of three "extreme" types of communities and one "other" type of community. Schools were placed into these categories on the basis of information about the type of community, the size of its population (as of the 1980 Census), and an occupational profile of residents provided by school principals before the assessment. The principals completed estimates of the percentage of students whose parents fit into each of six occupational categories. For those schools where the principal or his or her designate was unable or unwilling to answer the question on the occupational profile of parents, the type of community category was assigned as "missing."

The definitions of these "extreme" categories were determined using data from the 1992 national assessment. The categories are formed so that, approximately, an estimated 10 percent of the student population nationally at each grade level attend schools in each of the three "extreme" community types. These same criteria were then applied on a school-by-school basis to the schools that participated in the state assessments, to determine the type of community classification for each. The procedure for establishing these "extreme" classes using the national data has been similar throughout NAEP's history. This procedure is described in the sampling and weighting reports for the 1986, 1988, and 1990 national assessments (see Burke, Braden, Hansen, Lago, & Tepping, 1987; Rust, Bethel, Burke, & Hansen, 1990; and Rust, Burke, Fahimi, & Wallace, 1992). The type of community categories are as follows:

1 - Extreme Rural: Students in this group live outside metropolitan statistical areas, live in areas with a population below 10,000, and attend schools where many of the students' parents are farmers or farm workers.

2 - Disadvantaged Urban: Students in this group live in metropolitan statistical areas and attend schools where a high proportion of the students' parents are on welfare or are not regularly employed.

3 - Advantaged Urban: Students in this group live in metropolitan statistical areas and attend schools where a high proportion of the students' parents are in professional or managerial positions.

4 - Other: Students in this category attend schools in areas other than those defined as advantaged urban, disadvantaged urban, or extreme rural.

PARED (Parents' education level)

The variable PARED is derived from responses to two questions, B003501 and B003601, in the student demographic questionnaire. Students were asked to indicate the extent of their mother's education (B003501—How far in high school did your mother go?) by choosing one of the following:

○ She did not finish high school.
○ She graduated from high school.
○ She had some education after high school.
○ She graduated from college.
○ I don't know.

Students were asked to provide the same information about the extent of their father's education (B003601—How far in high school did your father go?) by choosing one of the following:

○ He did not finish high school.
○ He graduated from high school.
○ He had some education after high school.
○ He graduated from college.
○ I don't know.

The information was combined into one parental education reporting category (PARED) as follows: If a student indicated the extent of education for only one parent, that level was included in the data. If a student indicated the extent of education for both parents, the higher of the two levels was included in the data. For students who did not know the level of education for both parents or did not know the level of education for one parent and did not respond for the other, the parental education level was classified as unknown. If the student did not respond for both parents, the student was recorded as having provided no response.

REGION (Region of the country)

States were grouped into four geographical regions—Northeast, Southeast, Central, and West—as shown in Table E-1. All 50 states and the District of Columbia are listed, with the participants in the Trial State Assessment shown in italic type. Territories were not assigned to a region. The part of Virginia that is included in the Washington, DC, metropolitan statistical area is included in the Northeast region; the remainder of the state is included in the Southeast region.

Table E-1
NAEP Geographic Regions

NORTHEAST	SOUTHEAST	CENTRAL	WEST
Connecticut	*Alabama*	Illinois	Alaska
Delaware	*Arkansas*	*Indiana*	*Arizona*
District of Columbia	*Florida*	*Iowa*	*California*
Maine	*Georgia*	Kansas	*Colorado*
Maryland	*Kentucky*	*Michigan*	*Hawaii*
Massachusetts	*Louisiana*	*Minnesota*	*Idaho*
New Hampshire	*Mississippi*	*Missouri*	Montana
New Jersey	*North Carolina*	*Nebraska*	Nevada
New York	*South Carolina*	*North Dakota*	*New Mexico*
Pennsylvania	*Tennessee*	*Ohio*	*Oklahoma*
Rhode Island	*Virginia*	South Dakota	Oregon
Vermont	*West Virginia*	*Wisconsin*	*Texas*
Virginia			*Utah*
			Washington
			Wyoming

MODAGE (Modal age)

The modal age (the age of most of the students in the grade sample) for the fourth grade students is age 9. The modal age of the eighth-grade students is age 13. A value of 1 for MODAGE indicates that the student is younger than the modal age; a value of 2 indicates that the student is of the modal age; a value of 3 indicates that the student is older than the modal age.

VARIABLES DERIVED FROM THE STUDENT, TEACHER, AND SCHOOL QUESTIONNAIRES

Several variables were formed from the systematic combination of response values for one or more items from either the student demographic questionnaire, the student mathematics background questionnaire, the teacher questionnaire, or the school questionnaire.

HOMEEN2 (Home environment—Articles [of 4] in the home)

The variable HOMEEN2 was created from the responses to student demographic items B000901 (Does your family get a newspaper regularly?), B000903 (Is there an encyclopedia in your home?), B000904 (Are there more than 25 books in your home?), and B000905 (Does your family get any magazines regularly?). The values for this variable were derived as follows:

1 0-2 types	The student responded to at least two items and answered Yes to two or fewer.
2 3 types	The student answered Yes to three items.
3 4 types	The student answered Yes to four items.
8 Omitted	The student answered fewer than two items.

SINGLEP (How many parents live at home)

SINGLEP was created from items B005601 (Does either your mother or your stepmother live at home with you?) and B005701 (Does either your father or your stepfather live at home with you?). The values for SINGLEP were derived as follows:

1 2 parents at home	The student answered Yes to both items.
2 1 parent at home	The student answered Yes to B005601 and No to B005701, or Yes to B005701 and No to B005601.
3 Neither at home	The student answered No to both items.
8 Omitted	The student did not respond to or filled in more than one oval for one or both items.

PERCMAT (Eighth-grade students' perception of mathematics based on 5 questions)
PERMATB (Eighth-grade students' perception of mathematics based on 8 questions)

Eight of the questions in the eighth-grade mathematics background questionnaire asked students about their perceptions of mathematics (How much do you agree with each of the following statements?):

BEST COPY AVAILABLE

M810701	I like mathematics.
M810702	Almost all people use mathematics in their jobs.
M810703	I am good in mathematics.
M810704	Mathematics is more for boys than for girls.
M810705	Mathematics is useful for solving everyday problems.
M810706	If I had a choice, I would not study any more mathematics.
M810707	I understand most of what goes on in mathematics class.
M810708	Learning mathematics is mostly memorizing facts.

PERCMAT was created from the first five of these questions (M810701 through M810705); PERMATB was created from all eight. For each item, the student could respond as follows:

1. Strongly agree
2. Agree
3. Undecided
4. Disagree
5. Strongly disagree

To derive these variables, first the values for three items (M810704, M810706, M810708) were reversed (e.g., "strongly disagree" became 1). Then, for each of the eight items, values 3, 4, and 5 were combined to create one value (new value 3). PERCMAT was determined by adding the values for the first five items and dividing by five to obtain a mean; PERMATB was determined by adding the values for all eight items and dividing by eight to obtain a mean. The mean was then recoded as follows:

1 - 1.67 = 1 Strongly agree
1.68 - 2.33 = 2 Agree
2.34 - 3 = 3 Undecided, disagree, or strongly disagree

The student had to answer at least one of the applicable items to get a value for PERCMAT or PERMATB.

PERCMA2 (Fourth-grade students' perception of mathematics based on 5 questions)
PERMA2B (Fourth-grade students' perception of mathematics based on 8 questions)

Eight of the questions in the fourth-grade mathematics background questionnaire asked students about their perceptions of mathematics (How much do you agree with each of the following statements?):

M811101	I like mathematics.
M811102	Almost all people use mathematics in their jobs.
M811103	I am good in mathematics.
M811104	Mathematics is more for boys than for girls.
M811105	Mathematics is useful for solving everyday problems.
M811106	I understand most of what goes on in mathematics class.

M811107 Learning mathematics is mostly memorizing facts.
M811108 If I had a choice, I would not study any more mathematics.

PERCMA2 was created from the first five of these questions (M811101 through M811105); PERMA2B was created from all eight. For each item, the student could respond as follows:

1. Agree
2. Undecided
3. Disagree

To derive these variables, first the values for three items (M811104, M811107, M811108) were reversed ("disagree" became 1 and "agree" became 3). Then, for each of the eight items, values 2 and 3 were combined to create one value (new value 2). PERCMA2 was determined by adding the values for the first five items and dividing by five to obtain a mean; PERMA2B was determined by adding the values for all eight items and dividing by eight to obtain a mean. The mean was then recoded as follows:

1 - 1.50 = 1 Agree
1.51 - 2 = 2 Undecided or disagree

The student had to answer at least one of the applicable items to get a value for PERCMA2 or PERMA2B.

TCERTIF (Type of teaching certificate)

Items T040501 through T040505 (Do you have teaching certification in any of the following areas that is recognized by the state in which you teach?) in the teacher questionnaire were combined to produce TCERTIF. The following rules were used to determine the three values of TCERTIF.

1 Mathematics	The teacher responded Yes to T040504 (middle/junior high school or secondary mathematics)
2 Education	The teacher responded Yes to T040501 (elementary or middle/junior high school education [general]) and No to T040504
3 Other	Any other response

TUNDMJB (Undergraduate major)

Items T040701 through T040705 in the teacher questionnaire (What were your undergraduate major fields of study?) were used to determine TUNDMJB as follows:

1 Mathematics	The teacher responded Yes to T040703 (mathematics)

2 Mathematics education	The teacher responded Yes to T040704 (mathematics education) and No to T040703
3 Education	The teacher responded Yes to T040701 (education) and No to T040703 and T040704
4 Other	Any other response

TGRDMJB (Graduate major)

Items T040801 through T040806 in the teacher questionnaire (What were your graduate major fields of study?) were used to determine TGRDMJB as follows:

1 Mathematics	The teacher responded Yes to T040803 (mathematics)
2 Mathematics education	The teacher responded Yes to T040804 (mathematics education) and No to T040803
3 Education	The teacher responded Yes to T040801 (education) and No to T040803 and T040804
4 Other	Any other response

TMATHEX (Exposure in areas of mathematics)

Items T041602 through T041607 in the teacher questionnaire were used to determine teachers' exposure to the mathematics topics of number systems and numeration, measurement, geometry, probability/statistics, abstract/linear algebra, and calculus. For each area, teachers were asked to indicated up to four levels of exposure that applied to them: one or more college courses, part of a course, in-service training, and little or no exposure.

TMATHEX was derived by summing the areas for which teachers responded "one or more college courses," then categorizing the responses into three levels:

1 Five to six areas
2 Three to four areas
3 Zero to two areas

VARIABLES DERIVED FROM MATHEMATICS ITEMS

CALCUSE (Calculator-usage index)

In each calculator block, items were classified as calculator-suitable (items for which use of a calculator was either required or not inappropriate), and calculator-unsuitable (items for which use of calculator was inappropriate). For each item in a calculator block, students were asked to indicate whether or not they used a calculator in answering the items.

The 1992 examinees who were administered at least one of the calculator blocks were classified into two groups—"high" and "other." The "high" group consisted of those examinees who indicated that they had used a calculator for 65 percent or more of the calculator-suitable items that they attempted and had used a calculator on no more than one of the calculator-unsuitable items that they attempted. The "other" group consisted of everyone else. For the purpose of assigning students to the categories of this variable, interest was restricted to the set of items for which a student had indicated whether or not he or she had used a calculator. Items for which a student failed to indicate this were excluded from the calculation of percentages.

NORMIT (Normit Gaussian score)
NORMITE (Normit Gaussian score, estimation sample)
SCHMATH (School-level mean Gaussian score)

The normit score is a student-level Gaussian score based on the inverse normal transformation of the mid-percentile rank of a student's number-correct booklet score within that booklet. The normit scores were used to decide collapsing of variables, finalize conditioning coding, and check the results of scaling.

Each student in the Trial State Assessment in mathematics had two scores—one for the main assessment booklet (NORMIT) and one for the estimation booklet (NORMITE). The number-correct is based on the number of dichotomous items answered correctly plus the score obtained on extended constructed-response items. The mid-percentile rank is based on the formula:

$$\frac{CF(i)+CF(i-1)}{2N}$$

where CF(i) is the cumulative frequency at i items correct and N is the total sample size. If i = 0 then

$$\frac{CF(0)+\frac{CF(1)}{2}}{2N}$$

A school-level normit, SCHMATH was also created; this was the mean normit across all main assessment mathematics booklets (i.e., not including the estimation booklets) administered in a school.

VARIABLES RELATED TO PROFICIENCY SCALING

Proficiency Score Variables

Item response theory (IRT) was used to estimate average mathematics proficiency for each state and for various subpopulations, based on students' performance on the set of mathematics items they received. IRT provides a common scale on which performance can be reported for the nation, state, and subpopulations, even when all students do not answer the same set of questions. This common scale makes it possible to report on relationships between students' characteristics (based on their responses to the background questions) and their overall performance in the assessment.

A scale ranging from 0 to 500 was created to report performance for each of the five mathematics content areas: Numbers and Operations; Measurement; Geometry; Data Analysis, Statistics, and Probability; and Algebra and Functions. Each content-area scale was based on the distribution of student performance across all three grades assessed in the 1992 national assessment (grades 4, 8, and 12) and had a mean of 250 and a standard deviation of 50. A composite scale was created as an overall measure of students' mathematics proficiency. The composite scale was a weighted average of the five content area scales, where the weight for each content area was proportional to the relative importance assigned to the content area as specified in the mathematics objectives. An additional scale was created for the items designed to measure estimation abilities. Although the items comprising each scale were identical to those used for the national program, the item parameters for the Trial State Assessment scales were estimated from the combined data from all jurisdictions participating in the Trial State Assessment.

Scale proficiency estimates were obtained for all students assessed in the Trial State Assessment. The NAEP methods use random draws ("plausible values") from estimated proficiency distributions to compute population statistics. Plausible values are not optimal estimates of individual proficiency; instead, they serve as intermediate values to be used in estimating population characteristics. Chapter 8 provides further details on the computation and use of plausible values.

The proficiency score (plausible value) variables are provided on the student data files for each of the scales and are named as shown in Table E-2.

Table E-2
Scaling Variables for the 1992 Trial State Assessment Samples

Mathematics Scale	Data Variables
Numbers and Operations	MRPS11 to MRPS15
Measurement	MRPS21 to MRPS25
Geometry	MRPS31 to MRPS35
Data Analysis, Statistics, and Probability	MRPS41 to MRPS45
Algebra and Functions	MRPS51 to MRPS55
Composite	MRPCM1 to MRPCM5
Estimation	MRPES1 to MRPES5

SMEANM (School mean score)
SNSCHM (Number of schools ranked)
SRANKM (School rank)
SRNK3M (Top, middle, bottom third)

At each grade level, a mean mathematics composite score based on the values from the scaling variable MRPCM1 was calculated for each school included in that grade level's assessment using the students' sampling weights. The schools were then ordered from highest to lowest mean score within a jurisdiction—the school with the highest mean score was given a ranking of 1 and the school with the lowest mean score was given a ranking equal to the number of schools in the jurisdiction. Values for school rank are found in the variable SRANKM. The number of schools ranked (i.e., the number of schools in the jurisdiction with assessed students) is found in the variable SNSCHM.

These variables were later used in partitioning the schools within the national public-school comparison sample and the schools within each state into three groups based on their ranking (highest third, middle third, and lowest third). The data from the partitioning are found in the variable SRNK3M.

APPENDIX F

THE NAEP SCALE ANCHORING PROCESS FOR THE 1992 MATHEMATICS ASSESSMENT

The NAEP Scale Anchoring Process for the 1992 Mathematics Assessment

Ina V.S. Mullis and Eugene G. Johnson

Educational Testing Service

Introduction

Beginning with the 1984 assessments, NAEP has generally reported students' subject area proficiency on 0-to-500 scales. These scales are used to report achievement for students at the various grades or ages assessed, including differences between performance from assessment to assessment for the nation and for various subpopulations of interest. To date, NAEP has used item response theory techniques to develop proficiency scales for reading, mathematics, science, U.S. history, and civics.

Although average proficiency is an efficient summary measure, some of the most interesting NAEP results are those based on performance differences for different points in the scale distributions. To provide an interpretation for both the average results (What does a 306 on the 0-to-500 scale actually mean?) and changes in performance distributions (What does it mean that fewer students are reaching level 250?), NAEP invented a scale anchoring process to describe the characteristics of student performance at various levels along the scales—typically, at levels 200, 250, 300, and 350. The descriptions of student performance are presented in the reports accompanied by the percentages of students performing at or above the various scale levels.

Scale anchoring is a way of attaching meaning to a scale. Traditionally, meaning has been attached to educational scales by norm-referencing, that is, by comparing students at a particular scale level to other students. In contrast, the NAEP scale anchoring is accomplished by describing what students at selected levels know and can do.

The mathematics composite scale was anchored in 1990. Since the 1992 mathematics scales were linked to the 1990 scales and shared a common framework and a pool of common items with the 1990 assessment, it was expected that the 1990 anchor descriptions would still be appropriate for 1992. However, the anchoring process was conducted again on the 1992 data to update the descriptions to reflect the newer item types included in the assessment, and to permit the selection of items to exemplify each of the levels. Consequently, the anchoring of the 1992 mathematics composite was viewed as an enhancement of the 1990 anchoring, and, as anticipated, the 1992 descriptions are very similar to the 1990 descriptions, with some variations.

In brief, NAEP's scale anchoring procedure for the 1992 mathematics assessment—like the 1990 assessment—was based on comparing item-level performance by students at four levels on the 0-to-500 overall mathematics proficiency scale—levels 200, 250, 300, and 350. These values (corresponding to standard deviation units of 50 from the overall mean in 1990 of 250) are far enough apart to be noticeably different but not so far apart as to be trivial. This analysis delineated four sets of anchor items that discriminated between adjacent performance levels on the scale. The four sets of empirically derived anchor items were studied by a panel of distinguished mathematics educators, who carefully considered and articulated the types of knowledge, skills, and reasoning abilities demonstrated by correct responses to the items in each set. The 16 panelists and the NAEP staff involved in the process worked first in two independent groups to develop descriptions using the 1990 descriptions and the 1992 anchor items. As might be expected, the two sets of descriptions were quite similar, but not identical. Thus, the panelists subsequently met as a whole to review both sets of descriptions and decide how best to present the combined view of the entire group. Anchoring results for the 1992 mathematics assessment are presented in several different reports. Each report provides the descriptions accompanied by some or all (depending on the report) of the anchor items available for public release. For each grade level at which the item was administered, each item is accompanied by its overall proportion correct (overall p-value) for the total population assessed and the p-values for each anchor level. The various steps in the procedure are detailed below.

The Scale Anchoring Analysis

NAEP's scale anchoring is grounded in an empirical process whereby the scaled assessment results are analyzed to delineate sets of items that discriminate between adjacent performance levels on the scale[1]. For the 1992 mathematics assessment, as in the 1990 assessment, the levels were 200, 250, 300, and 350. For these four levels, items were identified that were likely to be answered correctly by students performing at a particular level on the scale and much less likely to be answered correctly by students performing at the next lower level.

To provide a sufficient pool of respondents, in identifying anchor items, students at level 200 were defined as those whose estimated mathematics proficiency (as defined by their first composite plausible value) was between 187.5 and 212.5, students at 250 were defined as those with estimated proficiency between 237.5 and 262.5, those at 300 had estimated proficiencies between 287.5 and 312.5, and those at 350 between 337.5 and 362.5. In theory, proficiency levels above 350 or below 200 could have been defined; however, so few students in the assessment performed at the extreme ends of the scale that it was not possible to do so.

The 1992 mathematics scale anchoring analysis was based on the scaled composite proficiency results for fourth, eighth, and twelfth graders participating in the 1992 national assessment. As illustrated below, for each item in the NAEP assessment, ETS determined the weighted percentage and raw frequency (unweighted count) for students at each of the four scale levels correctly answering the item. This was done for each of the grade levels at which the item was administered, and for the grade levels combined, if the item was administered at more than one grade level. For example, regardless of the grade level, the data for each item were analyzed as shown in the following sample.

---

[1] A detailed discussion of the theoretical underpinnings of scale anchoring can be found in Beaton and Allen (1992).

Sample Scale Anchoring Results				
Scale point	200	250	300	350
Weighted p-value	0.49	0.85	0.96	0.98
Raw frequency	902	1555	1271	276

It should be noted that the percentages of students answering the item correctly at each of the four scale levels differ from the proportion of students scoring above each score level and from the overall p-value for the total sample at any one grade level.

Because the extended constructed-response items were scored on an ordered scale with 5 scoring levels (0 to 4), the above procedure, which relies on the notion of a correct or an incorrect response to an item, had to be generalized. To fit into the anchoring framework, each extended constructed response item was converted into 4 pseudoitems by dichotomization at each of the values 1 through 4. Thus, the first pseudoitem was coded 1 for scores 1, 2, 3, and 4, and coded 0 otherwise; the second pseudoitem was coded 1 for scores 2, 3, 4 and coded 0 otherwise; the third pseudoitem was coded 1 for scores 3 and 4 and coded 0 otherwise; and the fourth pseudoitem was coded 1 for a score of 4 and coded 0 otherwise. These pseudoitems were then analyzed in the same manner as the items scored correct/incorrect.

As described below, criteria were applied to the scale-level results and an analysis conducted to delineate the items that discriminated between scale levels. Because it was the lowest level being defined, level 200 did not have to be analyzed in terms of the next lower level, but only for the percentage of students at that level answering the item correctly. More specifically, for an item to anchor at level 200:

1) The p-value for students at level 200 had to be greater than or equal to 0.65, and

2) the calculation of the p-value at that level had to have been based on at least 100 students to ensure adequate stability of the estimate of the p-value.

As an example, the following results are for an item anchoring at level 200:

Level 200 Anchor Item Results				
Scale point	200	250	300	350
Weighted p-value	0.65	0.89	0.98	1.00
Raw frequency	116	706	510	23

For an item to anchor at the remaining levels, additional criteria had to be met. For example, to anchor at level 250:

1) The p-value for students at level 250 had to be greater than or equal to 0.65;

2) the p-value for students at level 200 had to be less than or equal to 0.50;

3) the difference between the two p-values had to be at least 0.30; and

4) the calculations of the p-values at both levels 200 and 250 had to have been based on at least 100 students.

The following data set illustrates the results for a level 250 anchor item:

Level 250 Anchor Item Results

Scale point	200	250	300	350
Weighted p-value	0.38	0.75	0.89	0.98
Raw frequency	247	569	509	83

The principles used for level 250 were also used to identify anchor items at levels 300 and 350. For example, the following results were obtained for an item anchoring at level 300:

Level 300 Anchor Item Results

Scale point	200	250	300	350
Weighted p-value	0.11	0.28	0.83	1.00
Raw frequency	134	670	512	52

The results below are for an item anchoring at level 350:

Level 350 Anchor Item Results

Scale point	200	250	300	350
Weighted p-value	0.00	0.22	0.37	0.94
Raw frequency	50	324	585	241

In summary, for any given anchor item, 1) students at the item's anchor level are likely to answer the item correctly ($p_1 \geq .65$); students at the next lower level are somewhat unlikely to answer the item correctly ($p_2 \leq .50$); and students at the next lower level are less likely than students at the anchor level to answer the item correctly ($p_1 - p_2 \geq .30$). Collectively, as identified through this procedure, the 1992 NAEP mathematics items at each anchor level represented advances in students' understandings from one level to the next—mathematical areas where students at that level were more likely to answer items correctly than were students at the next lower level.

Preparing for the Mathematics Item Anchoring Panel Meeting

The analysis procedures described above yielded 22 items that anchored at level 200, 45 items at level 250, 59 items at level 300, and 43 items at level 350. Additionally, to provide some information for cross-referencing purposes, items that "almost anchored" were also identified. While these items did not satisfy the anchoring criteria, they did satisfy the following relaxed criteria: The p-value for students at the anchor level was at least .60, the p-value at the next lower level was no more than .55, the difference between the two p-values was at least .27, and the calculations of the p-values at both levels was based on at least 20 students. This procedure yielded additional items at each score point (level 200—8 items, level 250—27 items, level 300—29 items, level 350—34 items) that could be used for further context in developing descriptions. Of the 432 items included in the process, 149 (34%) anchored and 98 (23%) almost anchored. Table F-1 provides a breakdown of the number of anchored and almost anchored items by content area and by grade.

In preparation for use by the scale anchoring panelists, the items were placed in notebooks by section in the following order: anchored at 200, almost anchored at 200, anchored at 250, almost anchored at 250, anchored at 300, etc. Again, for further cross-referencing purposes, the remaining items in the assessment were also included in the notebook under the "did not anchor" heading. Each item was accompanied by its scoring guide (for constructed response items) and by the full anchoring documentation; that is, the anchoring information for each grade level at which an item was administered, the anchoring information across grades, the p-value for the total population of respondents at each grade level, and the mathematics content-area and process classifications for the items.

As described in *Mathematics Objectives, 1990 Assessment* (NAEP, 1988), which was also the framework for the 1992 assessment, the mathematics assessment was designed to measure five content areas, each with three ability levels. To ensure that the anchoring performance descriptions tied back to the assessment specifications, within anchor level sections, the items in the notebooks were sorted by the five content areas—Numbers and Operations; Measurement; Geometry; Data Analysis, Statistics, and Probability; and Algebra and Functions. Within content area, the items were sorted by ability level—procedural knowledge, conceptual understanding, and problem solving.

The Scale Anchoring Panel

Sixteen mathematics educators were invited to participate in the anchoring process. They represented teachers at the various grade levels involved, state mathematics supervisors from several of the states participating in the Trial State Assessment (including Washington, DC), large-city mathematics curriculum coordinators, and college mathematics professors and researchers. The group was also balanced by region of the country, race/ethnicity, and gender. (See Figure F-1 for a list of the participants.)

Table F-1
Counts of Items Anchoring and Almost Anchoring by Content Area and Grade

Content Area	DNA*	200	200A**	250	250A	300	300A	350	350A
GRADE 4:									
Num & Op	17	10	4	18	9	6	5	0	0
Measurement	6	7	1	5	6	4	2	0	0
Geometry	8	3	2	5	4	5	3	0	0
Data Analysis	9	0	0	10	1	2	1	0	0
Algebra & Fcns	8	1	0	6	0	1	2	0	0
GRADE 8:									
Num & Op	21	3	3	10	4	12	5	0	6
Measurement	6	1	2	4	4	7	1	6	4
Geometry	10	1	1	3	3	11	3	2	5
Data Analysis	16	0	0	3	4	7	3	1	3
Algebra & Fcns	8	2	0	3	1	9	2	2	5
GRADE 12:									
Num & Op	16	1	1	4	1	7	2	8	3
Measurement	11	1	0	1	0	3	1	8	3
Geometry	17	0	1	0	1	6	7	8	1
Data Analysis	21	0	0	1	1	7	0	2	0
Algebra & Fcns	29	1	0	0	1	5	1	17	5
ACROSS GRADES:									
Num & Op	41	10	4	18	11	14	10	8	8
Measurement	19	7	2	6	7	9	3	8	7
Geometry	32	3	2	5	4	14	9	8	6
Data Analysis	35	0	0	10	4	11	3	2	3
Algebra & Fcns	38	2	0	6	1	11	4	17	10

* Did not anchor
** Almost anchored

Figure F-1
Mathematics Scale Anchoring Panel

Charles Allen	Michigan Department of Education Lansing, Michigan
Thomas Carpenter	University of Wisconsin Madison, Wisconsin
John Dossey	Illinois State University Normal, Illinois
Edward Dubinsky	Purdue University West Lafayette, Indiana
David Farber	Voorhees High School Glen Gardener, New Jersey
Deborah Haimo	University of Missouri St. Louis, Missouri
Linda Kolnowski	Detroit Public Schools Detroit, Michigan
Gordon Lewis	Washington, DC, Public Schools Washington, DC
Mary Lindquist	Columbus College Columbus, Georgia
Donna Long	Indiana Department of Education Indianapolis, Indiana
Wendell Meeks	Illinois State Department of Education Springfield, Illinois
Tej Pandey	California Department of Education Sacramento, California
Mattye Pollard-Cole	Colorado State Department of Education Denver, Colorado
Diane Shaffer	Rhode Island Department of Education Providence, Rhode Island
Jocelyn Walton	Plainfield High School Plainfield, New Jersey
Charles Watson	Arkansas Department of Education Little Rock, Arkansas

The Process for Developing the Descriptions

The three-day anchoring meeting began on the morning of the first day, during which time panelists were thoroughly briefed in the anchoring process and given their assignment: With the objectives for the 1992 mathematics assessment and the anchor descriptions for the 1990 assessment as a reference, use the information in the anchor item notebooks to describe the mathematical knowledge, understandings, and problem-solving abilities demonstrated by the students at each anchor level in each of the five content areas. Based on the items anchoring at each anchor level (cross-referenced with "almost anchored" and "did not anchor" items), the panelists were asked to draft a description of achievement at each level in one-half page or less.

The meeting was structured so that the remainder of the first day and the entire second day could be devoted to the panelists working with staff in two independent groups to accomplish this task. In each of the independent groups, panelists and staff worked together to analyze the knowledge, skills, and reasoning abilities required by each item. Lists were developed portraying these for each mathematics content area at each anchor level. Based on these question-by-question analyses, which were prominently displayed around the room on poster paper, each group of panelists then drafted a description of performance for each anchor level. The two sets of draft descriptions appear as Figure F-2.

On the third day, the panelists and staff met as a whole to combine the two independently derived sets of descriptions. They also worked on developing short "titles" or descriptors for each category, and selecting example items to accompany the anchor level descriptions.

Both groups agreed that the two drafts were very similar and that with some final review and editing either set would have appropriately described the anchor item information. However, they did like the benefit of the cross-validation process and the fact that more people were able to participate in the process. As the group worked through the two descriptions, they identified preferences for some parts of each of the descriptions, resolved some issues, and made some formatting decisions. The combined view was checked by staff against the anchoring data, edited, and sent to the panelists for final review. The final draft of the descriptions is presented in Figure F-3.

Reporting the Anchor Item Results

Anchoring results are presented in three reports: the *NAEP 1992 Mathematics Report Card for the Nation and the States*, *Interpreting NAEP Scales*, and the *Data Compendium from the NAEP 1992 Mathematics Assessment for the Nation and the States*. In the first two reports, the anchor descriptions are supported by several items anchoring at each level. The *Data Compendium* includes all anchor items and almost anchor items available for public release. Each anchor item in the reports is, for each grade level at which it was assessed, accompanied by the overall percentage of success on the item as well as the anchor level information for each grade at which it was assessed. This is designed to reduce confusion between the percentages of success on the individual anchor items illustrating particular levels on the scale and the percentage of students who perform at or above each scale level.

Figure F-2
Draft Descriptions Prepared Independently by the Two Groups of Panelists

GROUP A

Level 200

Students at this level have a basic understanding of the whole numbers and their operations. They are able to write number names and select the largest four-digit number from a list. They can add and subtract most whole numbers without a calculator. They can add, subtract, multiply, and divide whole numbers with a calculator. In situations involving money they can round a number to the nearest dollar. These students can identify the solution to a one-step word problem.

These students can select appropriate instruments to measure length, weight, and temperature and can identify appropriate units to measure length. In geometry, these students can identify common shapes in two- and three-dimensions as well as select the results of flipping and turning shapes in the plane. They have a very elementary understanding of symmetry. In algebra, they are capable of extending a simple sequential symbol pattern.

Level 250

Students at this level have a solid understanding of whole number operations which allows them to translate between situations and mathematical representations for those settings. They can solve one- and simple two-step problems involving whole numbers, including interpreting the meaning of the resulting number. Their sense of whole numbers and their use extends to knowing when to estimate and what information in a problem situation may be extraneous.

GROUP B

Level 200

The students at this level function in the domain of whole numbers. They can add and subtract whole numbers, and when a calculator is available, they can multiply and divide. They can select the largest whole number from a set of numbers in the thousands, and can match the verbal and symbolic names for a number.

Length and weight are familiar attributes to these students in that they can select appropriate instruments and units to measure these attributes. They can recognize some basic properties of two-dimensional geometric figures as well as the names of standard examples of these figures. They can extend simple visual patterns.

Level 250

When presented with a problem situation, students have some understanding of the problem, they can identify extraneous information and have some knowledge of when to use computational estimation.

Students at this level have an understanding of addition, subtraction, multiplication, and division with whole numbers. They can solve routine one-step multiplication and division problems involving remainders and two-step addition and subtraction problems. They can round whole numbers

BEST COPY AVAILABLE

Figure F-2 (continued)
Draft Descriptions Prepared Independently by the Two Groups of Panelists

GROUP A

In measurement they can read scales on instruments, use a ruler to measure length in centimeters, and perform simple conversions of units in a system. They can use measurement scales in the solution of elementary word problems. Students at this level have extended abilities to deal with common planar shapes, including seeing them embedded in other figures or using them to dissect other figures. This emerging spatial sense allows them to visualize a cube and select counter-examples to elementary generalizations about the properties of a figure.

In data, these students are able to read, construct, and interpret data represented in tables, bar graphs, circle graphs and pictographs, including one- and two-step problems based on such data. They have an elementary understanding of relative frequency probability and related simple expectancy settings.

In the algebra dimension students at this level are able to read a number line and extend values to intermediate points. Their understanding also allows the extensions of simple arithmetic progression patterns in an applied setting.

Level 300

Students at this level are able to answer simple questions or solve simple problems involving fractions and decimals. They are able to both identify and create examples that illustrate equivalences of fractions and decimals, including locating the positions of such numbers on a number line. They can solve increasingly complex multi-step problems. Their understanding of percent includes the ability to calculate the effect of a

GROUP B

and solve simple word problems involving place value, estimation, and multiples.

These students can use a ruler to measure length and have some understanding of area and perimeter. They can solve simple problems using readings from instruments. They demonstrate a knowledge of properties of triangles, squares, rectangles, circles, and cubes. They can solve problems that require visualizing, drawing or manipulating simple geometric shapes. They can complete bar graphs and pictographs, as well as use information from graphs or tables to solve simple problems. They can recognize simple number patterns, are beginning to deal informally with the idea of a variable, and have some knowledge of simple probability.

Level 300

Students at this level can use various strategies and explain their reasoning in a variety of problem solving situations. They are able to solve problems involving not only whole numbers but with decimals and fractions. They can represent, compare, and find equivalent fraction, and use these concepts in solving routine problems. They can find a percent of a number and use this skill in simple problems. Multiplication and division of whole

Figure P-2 (continued)
Draft Descriptions Prepared Independently by the Two Groups of Panelists

GROUP A

percent increase on a total amount. With integers they are able to find simple products. At this level there is the emergence of some understanding of the number theory ideas of multiple and divisor.

In measurement these students can find areas of squares and rectangles and know the relationship between the perimeter of a square and the length of its side. They are able to give the measurement of the length of an object, use rulers with some flexibility, and use the scale on a map to approximate distance. Given a formula, they can substitute measures to get a numerical value.

In geometry, students at this level have a basic understanding of the properties of squares, rectangles, and parallelograms and are able to use those properties to identify necessary conditions and make some elementary indirect measurements. They are capable of finding the length of a missing side of a triangle in a simple similarity setting. They know that the sum of the measures of the angles of a triangle equal 180° and are able to use this property in simple problem settings. Using manipulatives they can combine shapes to represent a specified shape condition. Their spatial sense has increased to include the ability to visualize a cube in either three-space or its planar net arrangement.

In data analysis these students can draw data from a table to make decisions or, given additional data, insert new data in an existing table. They are able to draw data from circle and line graphs and compute with that data to answer questions or describe when an event occurred. They also have an understanding of bias in a sample. Their knowledge of probability is still rooted in relative frequencies in simple simulation situations. Students at this level can list the elements in a simple sample

GROUP B

and rational numbers have developed to the extent that students can use all four operations in multistep problems.

At this level student can read and use instruments in more complex situations. They can find areas of rectangles, recognize relationships among common units of measure, and use proportional relationships to solve routine problems involving similar triangles and scale drawings. In geometry, they have knowledge of definitions and properties of simple geometric figures in the plane. They can visualize composition and decomposition of two- and three-dimensional figures.

These students can calculate averages, select and interpret data from a variety of graphs, list the possible arrangements in a sample space, find the probability of a simple event, and have a beginning understanding of sample bias. They can evaluate algebraic expressions and solve linear inequalities by substitution, and solve equations involving square roots.

Figure F-2 (continued)
Draft Descriptions Prepared Independently by the Two Groups of Panelists

GROUP A

space and list the permutations of three objects.

In algebra these students can graph points on coordinate axes, locate the missing coordinates for a corner of a square, and identify which ordered pairs satisfy a given linear equation. With inequalities they can shade the points on a number line satisfying a simple interval given as an inequality, and solve elementary linear inequalities involving whole numbers. Students at this level show the ability to evaluate simple expressions and solve linear literal equations.

Level 350

Students at this level have extended their knowledge of number and variable to relate and represent large numbers using exponents and scientific notation. They can use exponents in evaluating the decimal value of a number displayed in scientific notation on a calculator or in solving equations involving powers of numbers. Their knowledge of percent extends to evaluating situations involving variables and estimating percents. Other situations reflect their ability to use rates in solving two-step application problems. In number theory they show a solid understanding of even and odd numbers and their properties under computation.

In measurement these students can solve a variety of perimeter and area problems involving triangles, quadrilaterals, and circles. Their concept of surface area allows the solution of problems involving rectangular solids. In geometry they can solve for the length of missing segments in more complex similarity situations involving the use of basic geometric theorems. Their overall knowledge of planar geometric concepts and relations has

GROUP B

Level 350

In problem solving, students are able to use their reasoning and analytic abilities when they encounter new situations. They can use reasoning strategies, data, models, and relevant mathematics in solving problems. Students can judge the reasonableness and correctness of their solutions.

Students at this level can reason and estimate with percents. They can recognize scientific notation and find the decimal equivalent. They can apply their knowledge of area and perimeter of simple geometric figures to solve problems. They can find the circumferences of circles and the surface areas of solid figures. Students can apply the Pythagorean theorem to fund the hypotenuse of a right triangle. They are beginning to use rectangular coordinates in problem-solving situations and can apply geometric properties and relationships in solving problems.

Figure F-2 (continued)
Draft Descriptions Prepared Independently by the Two Groups of Panelists

GROUP A

extended to the coordinate plane and includes slope, distance, and some ideas of the rate of change in linear settings. Students at this level know and are able to apply the Pythagorean theorem in a variety of settings.

In data these students can interpret information supplied by a graph of a step function and calculate the mean (average) from a table of grouped data. In combinatorial problems these students can list the possible occurrences and examine them to solve problems.

In algebra and functions, besides the growth of coordinate geometry, these students have an extended understanding of an ability to use the properties of exponents in equation solving and computation. They can solve complex literal equations and systems of linear equations. With functions, they can evaluate a quadratic function for a given value, as well as find the value for a composite function. Graphically they can identify the zeros of a function and the graphical effect of taking the absolute value of a given function. Their knowledge of trigonometry includes the ability to find the trigonometric value associated with an angle in a right triangle, evaluate a functional value of an angle given in radian measure on the unit circle, and identify the value of a trigonometric expression using a basic trigonometric identity. Additionally these students show the ability to evaluate and represent complex patterns involving both numbers and expressions including variables.

GROUP B

The students can compute means from frequency tables and create a sample space to determine probabilities. Students can use exponents and evaluate expressions given in functional notation. They can identify an equation describing a linear relation provided in a table, solve literal equations and systems of two linear equations. They have some beginning knowledge of trigonometric relations. They can interpret a graph to determine the zeros of a function, read values in a step function, and transform a graph by applying the absolute value. They are able to recognize patterns in order to solve problems.

Figure F-3
Description of Mathematics Proficiency for
Four Anchor Levels on the NAEP Scale

Level 200	**Addition and Subtraction, and Simple Problem Solving with Whole Numbers**

Students at this level can identify solutions to one-step word problems, involving addition or subtraction. They can add and subtract whole numbers in most situations, and when a calculator is available, they can multiply and divide. They are able to select the largest whole number from a set of numbers in the thousands, and can match the verbal and symbolic names for numbers.

Students demonstrate familiarity with length and weight, by selecting appropriate instruments and units to measure these attributes. They are able to recognize some basic properties of two-dimensional geometric figures as well as the names of standard examples of these figures. They can recognize simple patterns.

Level 250	**Multiplication and Division, Simple Measurement, and Two-Step Problem Solving**

When presented with a problem situation, students at this level have some understanding of the problem, can identify extraneous information, and have some knowledge of when to use computational estimation. They have an understanding of addition, subtraction, multiplication, and division with whole numbers. They can solve simple two-step problems involving whole numbers. They are able to round whole numbers and solve simple word problems involving place value, estimation, and multiples.

Students can use a ruler to measure length in centimeters and have some understanding of area and perimeter. They can solve simple problems using readings from instruments. They demonstrate a knowledge of properties of triangles, squares, rectangles, circles, and cubes. They can solve problems that require visualizing, drawing or manipulating simple geometric shapes. They are able to complete bar graphs and pictographs, as well as use information from graphs or tables to solve simple problems. They can recognize simple number patterns, are beginning to deal informally with the idea of a variable, and have some knowledge of simple probability.

(continued)

Figure F-3 (continued)
Description of Mathematics Proficiency for
Four Anchor Levels on the NAEP Scale

Level 300	**Reasoning and Problem-Solving Involving Fractions, Decimals, Percents, and Elementary Concepts in Geometry, Statistics, and Algebra**

Students at this level can use various strategies and explain their reasoning in a variety of problem-solving situations. They are able to solve problems involving not only whole numbers but with decimals and fractions. They can represent and find equivalent fractions, and use these concepts in solving routine problems. They can find a percent of a number and use this skill in simple problems. Multiplication and division of whole numbers have developed to the extent that students can use all four operations in multistep problems.

Students can read and use instruments in more complex situations. They can find areas of rectangles, recognize relationships among common units of measure, and solve routine problems involving similar triangles and scale drawings. They have knowledge of definitions and properties of simple geometric figures in the plane. Their spatial sense includes the ability to visualize a cube in either three-space or its flattened form in a plane.

Students can calculate averages, select and interpret data from a variety of graphs, list the possible arrangements in a sample space, find the probability of a simple event, and have a beginning understanding of sample bias. They can use knowledge of relative frequencies in simple simulation situations. Students show the ability to evaluate simple expressions and solve linear equations. Students can graph points on coordinate axes, locate the missing coordinates for a corner of a square, and identify which ordered pairs satisfy a given linear equation.

Level 350	**Reasoning and Problem Solving Involving Geometric Relationships, Algebra, and Functions**

Students at this level can reason and estimate with percents. They can recognize scientific notation and find the decimal equivalent. They can apply their knowledge of area and perimeter of simple geometric figures to solve problems. They can find the circumferences of circles and the surface areas of solid figures. They can solve for the length of missing segments in more complex similarity situations. Students can apply the Pythagorean theorem to find the hypotenuse of a right triangle. They are beginning to use rectangular coordinates in problem-solving situations and can apply geometric properties and relationships in solving problems.

Students can compute means from frequency tables and create a sample space to determine probabilities, and read the graph of a step function. Students can use exponents and evaluate expressions given in functional notation. In number theory, they have an understanding of even and odd numbers and their properties. They can identify an equation describing a linear relation provided in a table, and solve literal equations and systems of two linear equations. They have some knowledge of trigonometric relations. These students can represent and interpret complex patterns and data using numbers, expressions, and graphs. Given the graph of a function they can identify its zeros and the effect on the graph of taking the absolute value of the function.

APPENDIX G

THE NAEP ACHIEVEMENT LEVEL SETTING PROCESS FOR THE 1992 MATHEMATICS ASSESSMENT

The NAEP Achievement Level Setting Process for the 1992 Mathematics Assessment

Mary Lyn Bourque

National Assessment Governing Board

Introduction

Since 1984 NAEP has reported the performance of students in the nation and for specific subpopulations on a 0-to-500 proficiency scale. The history and development of the scale and the anchoring procedure used to interpret specific points on that scale is described elsewhere in this report.

However, the 1988 legislation[1] created an independent board, the National Assessment Governing Board (NAGB), responsible for setting policy for the NAEP program. The Board has a statutory mandate to identify "appropriate achievement goals for each. . .grade in each subject area to be tested under the National Assessment." Consistent with this directive, and striving to achieve one of the primary mandates of the statute "to improve the form and use of NAEP results," the Board set performance standards (called achievement levels by NAGB) for the National Assessment in 1990 and again in 1992.

The 1990 trial, initiated in December 1989 with the dissemination of a draft policy statement (NAGB, 1989) and culminating 22 months later in the publication of the NAGB report, *The Levels of Mathematics Achievement* (Bourque & Garrison, 1991), consisted of two phases: the main study and a replication-validation study. Although there were slight differences between the two phases, there were many common elements. Both phases used a modified (iterative/empirical) Angoff (1971) procedure for arriving at the levels; both focused on estimating performance levels based on a review of the 1990 NAEP mathematics item pool; and both phases employed a set of policy definitions for Basic, Proficient, and Advanced (NAGB, 1990) as the criteria for making the item ratings. However, the 1990 process was evaluated by a number of different groups (see Hambleton & Bourque, 1991) who identified technical flaws in the 1990 process. These evaluations influenced NAGB's decision to set the levels again in 1992 and to not use the 1990 levels as benchmarks for progress toward the national goals during the coming decade. However, it is interesting to note that the 1990 and 1992 processes produced remarkably similar results.

In September 1991 NAGB contracted with American College Testing (ACT) to convene the panels of judges that would recommend the levels on the 1992 NAEP assessments in reading, writing, and mathematics. While the 1992 level-setting activities were not unlike those undertaken by NAGB in 1990, there were significant improvements made in the process for 1992. There was a concerted effort to bring greater technical expertise to the process: The contractor selected by NAGB has a national reputation for setting standards in a large number of certification and licensure exams; an internal and external advisory team monitored all the technical decisions made by the contractor throughout the process; and state assessment directors periodically provided their expertise and technical assistance at key stages in the project.

---

[1] Public Law 100-297. (1988). National assessment of educational progress improvement act (Article No. USC 1221). Washington, DC.

Setting achievement levels is a method for setting standards on the NAEP assessment that identifies what students should know and be able to do at various points along the proficiency scale. The initial policy definitions of the achievement levels were presented to panelists along with an illustrative framework for more in-depth development and operationalization of the levels. Panelists were asked to determine descriptions/definitions of the three levels from the specific framework developed for the NAEP assessment with respect to the content and skills to be assessed. The operationalized definitions were refined throughout the level-setting process, as well as validated with a supplementary group of judges subsequent to the level-setting meetings. Panelists were also asked to develop a list of illustrative tasks associated with each of the levels, after which sample items from the NAEP item pool were identified to exemplify the full range of performance of the intervals between levels. The emphasis in operationalizing the definitions and in identifying and selecting exemplar items and papers was to represent the full range of performance from the lower level to the next higher level. The details of the implementation procedures are outlined in the remainder of this appendix.

Preparing for the 1992 Mathematics Level-setting Meeting

It is important for the planning of any standard-setting effort to know how various process elements interact with each other. For example, panelists interact with pre-meeting materials, the meeting materials (i.e., the assessment questions, rating forms, rater feedback, and so forth), each other, and the project staff. All of these elements combine to promote or degrade what has been called intrajudge consistency and interjudge consensus (Friedman & Ho, 1990).

Previous research has conceptualized the effects of two major kinds of interaction: people interacting with text (Smith & Smith, 1988) and people interacting with each other (Curry, 1987; Fitzpatrick, 1989). To assess the effects of textual and social interaction and adjust the standard setting procedures accordingly, a pilot study was conducted as the first phase of the 1992 initiative.

Reading was chosen as the single content area to be pilot-tested since it combined all of the various features found in the other NAEP assessments, including multiple-choice, short constructed-response, and extended constructed-response items. The pilot study provided the opportunity to implement and evaluate all aspects of the operational plan—background materials, meeting materials, study design, meeting logistics, staff function, and participant function.

The overall pilot effort was quite successful. The level-setting process worked well and the pilot allowed the contractor to make improvements in the design before implementation activities began. For example, schedule changes were made that allowed the panelists more time to operationalize the policy definitions before beginning the item-rating task. Also, the feedback mechanisms used to inform panelists about interjudge and intrajudge consistency data were improved for clarity and utility to the entire process.

The Mathematics Level-setting Panel

Sixty-nine panelists representing 32 jurisdictions (31 states and the District of Columbia) from the 424 nominees were invited to participate in the level-setting process. They represented mathematics teachers at grades 4, 8, and 12, nonteacher educators, and members of the noneducator (general public) community. The group was balanced by gender, race/ethnicity, NAEP regions of the country, community type (low SES, not low SES), district size, and school type (public/private). One panelist was unable to attend due to a family emergency, resulting in 68 participants: 24 at grade 4 and 22 at grades 8 and 12.

Process for Developing the Achievement Levels

The four-and-one-half day session began with a brief overview of NAEP and NAGB, a presentation on the policy definitions of the achievement levels, a review of the NAEP mathematics assessment framework, and a discussion of factors that influence item difficulty. The purpose of the presentation was to focus panelists' attention on the mathematics framework and to emphasize the fact that panelists' work was directly related to the NAEP assessment, not to the whole domain of mathematics.

All panelists completed and self-scored an appropriate grade-level form of the NAEP assessment. The purpose of this exercise was to familiarize panelists with the test content and scoring protocols before beginning to develop the preliminary operationalized descriptions of the three levels.

Working in small groups of five or six, panelists expanded and operationalized the policy definitions of Basic, Proficient, and Advanced in terms of specific mathematical skills, knowledge, and behaviors that were judged to be appropriate expectations for students in each grade, and were in accordance with the current mathematics assessment framework.

The policy definitions are as follows:

Basic	This level, below proficient, denotes partial mastery of the knowledge and skills that are fundamental for proficient work at each grade—4, 8, and 12.
Proficient	This central level represents solid academic performance for each grade tested—4, 8, and 12. Students reaching this level have demonstrated competency over challenging subject matter and are well prepared for the next level of schooling.

Advanced This higher level signifies superior performance beyond proficient grade-level mastery at grades 4, 8, and 12.

The small groups were allowed to brainstorm about what student performance *should* be, using the framework and their experience in completing the NAEP assessment as guides[2]. A comprehensive listing of grade-level descriptors was developed, and panelists were asked to identify the five that best described what students *should* be able to do at each of the levels. Those descriptors appearing with the greatest frequency were compiled into a discussion list for the grade-level groups. Additions, deletions, and modifications were made as a result of discussions, and the groups reached general agreement that the final list of descriptors represented what students *should* be able to do at each achievement level.

Panelists next received two hours of training in the Angoff method. Training was customized to reflect the unique item formats of the particular subject area assessment. Once a conceptual consensus was reached about the characteristics of *marginally* acceptable examinees at each of the three levels, practice items from the released pool were rated by the panelists according to the process defined in the contractor's plan. For multiple-choice and short constructed-response items, panelists were asked to rate each item for the expected probability of a correct response for a group of *marginally* acceptable examinees at the Basic, Proficient, and Advanced levels. For extended constructed-response items, panelists were asked to review 20 to 25 student response papers and select three papers, one for each achievement level, that typified *marginally* acceptable examinee performance for that level.

Following training in the Angoff method, the judges began the rating process, inspecting and rating each item in the pool for the expected probabilities of answering the item correctly at each level. Panelists completed three rounds of item ratings. For Round 1, panelists first answered the items in each section, then reviewed their answers using scoring keys and protocols. This process helped ensure that panelists would be thoroughly familiar with each item, including the foils and scoring rubrics, before rating the items. Panelists provided item ratings/paper selections for all three achievement levels, one item at a time, for all the items in a section, then proceeded to the next set of items, for which the process was repeated. During Round 1, panelists used their lists of descriptors and other training materials for guidance in the rating process.

Following Round 1, item response theory (IRT) was used to convert the rating results[3] for each rater to a latent ability scale represented by the Greek letter theta (θ). This θ scale was the same scale used to calibrate the NAEP items evaluated by each panelist. In order to provide meaningful feedback about item ratings, a special *relative scale* was constructed, which was a linear transformation of the theta scale having a mean of 75 and standard deviation of 15. Before Round 2 of the rating process, panelists were given interjudge consistency information using this relative scale. This information allowed panelists to see on the scale where their individual mean item ratings were, relative to the mean for the group and to the means for other panelists. Reasons for extreme mean ratings, including the possibility that some panelists misinterpreted the item rating task, were discussed briefly.

---

[2] The panelists also reviewed about half the item pool (the half they would not be rating later) so the descriptors could be further modified if appropriate.

[3] Because the IRT item parameters were not available for the polytomously scored (extended constructed-response) items, these items (five at grade 4, six each at grades 8 and 12) were not included in the following discussion of results.

Before Round 2, panelists were also given item difficulty data. This information was presented as the percentage of students who answered each item correctly during the actual NAEP administration, for items scored "correct" or "incorrect" (i.e., multiple-choice and short constructed-response items), and as the percentage of students receiving scores of 1, 2, 3, and 4 for the extended constructed-response items[4]. Panelists were told that this item difficulty information should be used as a reality check. For items on which item ratings differed substantially from the item difficulty value, panelists were asked to reexamine the item to determine if they had misinterpreted the item or misjudged its difficulty. Results of the data analysis, and panelists' own evaluations, indicated that the item difficulty information was perceived as very useful but had little impact on panelists' ratings.

For Round 2, panelists reviewed the same set of items they had rated in Round 1 and, using the interjudge consistency information, the item difficulty information, and the information provided prior to Round 1, they either confirmed their initial item ratings or adjusted their ratings to reflect the additional information. About one-third of Round 1 item ratings were adjusted during Round 2.

Following Round 2, panelists' ratings were reanalyzed and additional information was presented to panelists concerning intrajudge variability prior to Round 3. For each panelist, the intrajudge variability information consisted of those items that they had rated differently than items having similar difficulty, taking into consideration the panelist's aggregated item ratings. That is, the panelists' aggregated item ratings were converted to the theta (θ) scale. All items rated by the panelists were then analyzed in terms of the panelist's achievement level (θ) in comparison to actual student performance on the items. The observed item rating from each panelist was contrasted to an expected item rating. Those items with large differences between observed and expected ratings were identified. Panelists were given this information and asked to review each of these items and decide if their Round 2 ratings still accurately reflected their best judgments of the items. The intrajudge consistency data was to be used to flag items for reconsideration in the final round of rating.

For Round 3, panelists reviewed the same set of items they rated in Rounds 1 and 2 using both the new intrajudge variability information and the information made available during Rounds 1 and 2. In addition, panelists could discuss, within their small groups, ratings of specific items about which they were unsure. About 20 percent of the item ratings were adjusted during Round 3.

---

[4] The percentages presented to the raters summed to 100 percent, but this excluded the percentages—around 80 percent, in some cases—of students who wrote responses that were judged to be "off-task," those who "skipped" that question and continued beyond that question, and those who, apparently, "never reached" that question.

Process of Selecting Exemplar Items

Following the standard-setting meeting, a series of procedures was implemented to select exemplar items. First, expected and empirical p-values were computed for each item in the released item pool. Expected p-values were based on predicted performance at the cut-off score for each achievement level and empirical p-values were based on the average performance of all students responding to the item. Items that did not have expected p-values ≥ 0.51 for any of the levels were deleted from the item pool. Second, items were compared to the operationalized descriptions of the levels. Items that did not match the content of the descriptions were deleted from the item pool. Third, the remaining items were classified as possible Basic, Proficient, or Advanced exemplars based on content match. Fourth, the validation panel reviewed the items and recommended a set of items to serve as exemplars for the levels. The final set of items was reviewed and approved by NAGB at their May 1992 meeting. These procedures are described in detail below.

Using the standard-setting ratings, expected p-values were computed for each item at the cut point for each achievement level. The criteria described below were applied to the scale-level results and an analysis was conducted to delineate items that could serve as exemplars for each achievement level (Basic, Proficient, Advanced). More specifically, for an item to be chosen as a possible exemplar for the Basic achievement level:

1) The expected p-value for students at the cut point for the Basic level of achievement had to be greater than 0.51;

2) The content of the item had to match the content of the operationalized description of Basic; and

3) The empirical p-value for the item had to be higher than empirical p-values for items selected as exemplars for the Proficient level.

As an example:

Grade 4 Basic Level Item M022801			
Level	**Basic**	**Proficient**	**Advanced**
Scale point	211	248	280
Expected p-value	0.70	0.82	0.94
Empirical p-value = 0.52			

For an item to be chosen as a possible exemplar for the Proficient achievement level:

1) The expected p-value for students at the cut-off score for the Proficient level of achievement had to be greater than 0.51;

2) The content of the item had to match the content of the operationalized description of Proficient; and

3) The empirical p-value for the item had to be lower than empirical p-values for Basic exemplar items, but higher than student p-values for Advanced exemplar items.

As an example:

Grade 4 Proficient Level Item M022001			
Level	**Basic**	**Proficient**	**Advanced**
Scale point	211	248	280
Expected p-value	0.37	0.58	0.76
Empirical p-value = 0.35			

For an item to be chosen as a possible exemplar for the Advanced achievement level:

1) The expected p-value for students at the cut-point for the Advanced level of achievement had to be greater than 0.51;

2) The content of the item had to match the content of the operationalized description of Advanced; and

3) The empirical p-value for the item had to be lower than empirical p-values for Proficient exemplar items.

As an example:

Grade 4 Advanced Level Item M023101			
Level	**Basic**	**Proficient**	**Advanced**
Scale point	211	248	280
Expected p-value	0.29	0.43	0.61
Empirical p-value = 0.22			

The analysis procedures described above yielded 31 items as possible grade 4 exemplars, 43 items as possible grade 8 exemplars, and 37 items as possible grade 12 exemplars, as follows:

Possible Exemplar Items by Grade and Achievement Level			
Grade	**Basic**	**Proficient**	**Advanced**
4	9	14	8
8	23	15	5
12	14	16	7

For grade 4, the possible exemplars represented 49 percent of the released item pool. For grades 8 and 12, the possible exemplars represented 54 percent of the released item pool for each grade.

Process for Validating the Levels

Eighteen mathematics educators participated in the item selection and content validation process. Ten of the panelists were mathematics teachers who had participated in the original achievement levels-setting process and who had been identified as outstanding panelists by grade group facilitators during this meeting. The other eight panelists represented the National Council of Teachers of Mathematics, the Mathematical Sciences Education Board, and state-level mathematics curriculum supervisors. To the extent possible, the group was balanced by race/ethnicity, gender, community type, and region of the country.

The two-and-one-half day meeting began by briefing panelists on the purpose of the meeting. They first reviewed the operationalized descriptions of the achievement levels for consistency with the NAGB policy definitions of Basic, Proficient, and Advanced and with the NAEP *Mathematics Objectives*. Next, they reviewed the operationalized descriptions of the achievement levels for qualities such as within- and across-grade consistency, grade-level appropriateness, and utility for increasing the public's understanding of the NAEP mathematics results. Finally, working first in grade level (4, 8, and 12) groups of six panelists each, then as a whole group, panelists revised the operationalized descriptions to provide more within- and across-grade consistency and to align the language of the description more closely with the language of the NCTM *Standards*. Both the original descriptions and the revised descriptions are included later in this appendix.

On the third day, panelists again split into grade-level groups of six panelists each and reviewed the possible exemplar items. The task was to select a set of items, for each achievement level for their grade, that would best communicate to the public the levels of mathematics ability and the types of skills needed to perform in mathematics at that level.

After selecting sets of items for their grades, the three grade-level groups met as a whole group to review item selection. During this process, cross-grade items that had been selected as exemplars by two grade groups (three such items were selected by grade groups 4 and 8) were assigned to one grade by whole-group consensus. In addition, items were evaluated by the whole group for overall quality. Two items were rejected by the group during this process due to possible bias. This process yielded 14 items as recommended exemplars for grade 4, 11 items as recommended exemplars for grade 8, and 14 items as recommended exemplars for grade 12.

Mapping Panelists' Ratings to the NAEP Scales

The process of mapping panelists' ratings to the NAEP scales made significant use of *item response theory* (IRT). IRT provides statistically sophisticated methods for determining the expected performance of examinees on particular test items in terms of an appropriate measurement scale. The same measurement scale simultaneously describes the characteristics of the test items and the performance of the examinees. Once the item characteristics are set, it is possible to precisely determine how examinees are likely to perform on the test items at different points of the measurement scale.

The panelists' ratings of the NAEP test items were likewise linked, by definition, to the expected performance of examinees at the theoretical achievement level cut points. It was therefore feasible to use the IRT item characteristics to calculate the values on the measurement scale corresponding to each achievement level. This was done by averaging the item ratings over panelists for each achievement level and then simply using the item characteristics to find the corresponding achievement level cut points on the IRT measurement scale. This process was repeated for each of the NAEP content areas within each grade (4, 8, and 12).

In the final stage in the mapping process, the achievement level cut points on the IRT measurement scale were combined over content areas and rescaled to the NAEP score scale. Weighted averages of the achievement level cut points were computed. The weighting constants accounted for the measurement precision of the test items evaluated by the panelists, the proportion of items belonging to each NAEP content area, and the linear NAEP scale transformation. These weighted averages produced the final cut points for the Basic, Proficient, and Advanced achievement levels within each grade.

Figure G-1
Final Description of 1992 Mathematics Achievement Levels

GRADE 4

The NAEP content areas include: (1) numbers and operations; (2) measurement; (3) geometry; (4) data analysis, statistics, and probability; (5) algebra and functions. (Note: At the fourth-grade level, algebra and functions are treated in informal and exploratory ways, often through the study of patterns.) Skills are cumulative across levels—from Basic to Proficient to Advanced.

BASIC. Fourth-grade students performing at the **basic level** *should show some evidence of understanding the mathematical concepts and procedures in the five NAEP content areas.*

Specifically, fourth graders performing at the basic level should be able to estimate and use basic facts to perform simple computations with whole numbers; show some understanding of fractions and decimals; and solve simple real-world problems in all NAEP content areas. Students at this level should be able to use—though not always accurately—four-function calculators, rulers, and geometric shapes. Their written responses are often minimal and presented without supporting information.

PROFICIENT. Fourth-grade students performing at the **proficient level** *should consistently apply integrated procedural knowledge and conceptual understanding to problem solving in the five NAEP content areas.*

Specifically, fourth graders performing at the proficient level should be able to use whole numbers to estimate, compute, and determine whether results are reasonable. They should have a conceptual understanding of fractions and decimals; be able to solve real-world problems in all NAEP content areas; and use four-function calculators, rulers, and geometric shapes appropriately. Students performing at the proficient level should employ problem-solving strategies such as identifying and using appropriate information. Their written solutions should be organized and presented both with supporting information and explanations of how they were achieved.

ADVANCED. Fourth-grade students performing at the **advanced level** *should apply integrated procedural knowledge and conceptual understanding to complex and nonroutine real-world problem solving in the five NAEP content areas.*

Specifically, fourth graders performing at the advanced level should be able to solve complex and nonroutine real-world problems in all NAEP content areas. They should display mastery in the use of four-function calculators, rulers, and geometric shapes. These students are expected to draw logical conclusions and justify answers and solution processes by explaining why, as well as how, they were achieved. They should go beyond the obvious in their interpretations and be able to communicate their thoughts clearly and concisely.

Figure G-1 (continued)
Final Description of 1992 Mathematics Achievement Levels

GRADE 8

NAEP content areas: (1) numbers and operations; (2) measurement; (3) geometry; (4) data analysis, statistics, and probability; (5) algebra and functions. Skills are cumulative across all levels—from Basic to Proficient to Advanced.

BASIC. Eighth-grade students performing at the **basic level** *should exhibit evidence of conceptual and procedural understanding in the five NAEP content areas.* This level of performance signifies an understanding of arithmetic operations—including estimation—on whole numbers, decimals, fractions, and percents.

Eighth graders performing at the basic level should complete problems correctly with the help of structural prompts such as diagrams, charts, and graphs. They should be able to solve problems in all NAEP content areas through the appropriate selection and use of strategies and technological tools—including calculators, computers, and geometric shapes. Students at this level also should be able to use fundamental algebraic and informal geometric concepts in problem solving.

As they approach the proficient level, students at the basic level should be able to determine which of available data are necessary and sufficient for correct solutions and use them in problem solving. However, these eighth graders show limited skill in communicating mathematically.

PROFICIENT. Eighth-grade students performing at the **proficient level** *should apply mathematical concepts and procedures consistently to complex problems in the five NAEP content areas.*

They should be able to conjecture, defend their ideas, and give supporting examples. They should understand the connections between fractions, percents, decimals, and other mathematical topics such as algebra and functions. Students at this level are expected to have a thorough understanding of basic-level arithmetic operations—an understanding sufficient for problem solving in practical situations.

Quantity and spatial relationships in problem solving and reasoning should be familiar to them, and they should be able to convey underlying reasoning skills beyond the level of arithmetic. They should be able to compare and contrast mathematical ideas and generate their own examples. These students should make inferences from data and graphs; apply properties of informal geometry; and accurately use the tools of technology. Students at this level should understand the process of gathering and organizing data and be able to calculate, evaluate, and communicate results within the domain of statistics and probability.

ADVANCED. Eighth-grade students at the **advanced level** *should be able to reach beyond the recognition, identification, and application of mathematical rules in order to generalize and synthesize concepts and principles in the five NAEP content areas.*

They should be able to probe examples and counter examples in order to shape generalizations from which they can develop models. Eighth graders performing at the advanced level should use number sense and geometric awareness to consider the reasonableness of an answer. They are expected to use abstract thinking to create unique problem-solving techniques and explain the reasoning processes underlying their conclusions.

Figure G-1 (continued)
Final Description of 1992 Mathematics Achievement Levels

GRADE 12

NAEP content areas: (1) numbers and operations; (2) measurement; (3) geometry; (4) data analysis, statistics, and probability; (5) algebra and functions. Skills are cumulative across levels—from Basic to Proficient to Advanced.

BASIC. Twelfth-grade students at the **basic level** *should demonstrate procedural and conceptual knowledge in solving problems in the five NAEP content areas.*

They should be able to use estimation to verify solutions and determine the reasonableness of results as applied to real-world problems. They are expected to use algebraic and geometric reasoning strategies to solve problems. Twelfth graders performing at the basic level should recognize relationships presented in verbal, algebraic, tabular, and graphical forms; and demonstrate knowledge of geometric relationships and corresponding measurement skills.

Twelfth graders at the basic level should be able to apply statistical reasoning in the organization and display of data and in reading tables and graphs. They also should be able to generalize from patterns and examples in the areas of algebra, geometry, and statistics. At this level, they should use correct mathematical language and symbols to communicate mathematical relationships and reasoning processes; and use calculators appropriately to solve problems.

PROFICIENT. Twelfth-grade students at the **proficient level** *should consistently integrate mathematical concepts and procedures to the solutions of more complex problems in the five NAEP content areas.*

Twelfth graders performing at the proficient level should demonstrate an understanding of algebraic, statistical, and geometric and spatial reasoning. They should be able to perform algebraic operations involving polynomials; justify geometric relationships; and judge and defend the reasonableness of answers as applied to real-world situations. These students should be able to analyze and interpret data in tabular and graphical form; understand and use elements of the function concept in symbolic, graphical, and tabular form; and make conjectures, defend ideas, and give supporting examples.

ADVANCED. Twelfth-grade students at the **advanced level** *should consistently demonstrate the integration of procedural and conceptual knowledge and the synthesis of ideas in the five NAEP content areas.*

They should understand the function concept; and be able to compare and apply the numeric, algebraic, and graphical properties of functions. They should apply their knowledge of algebra, geometry, and statistics to solve problems in more advanced areas of continuous and discrete mathematics.

Twelfth graders performing at the advanced level should be able to formulate generalizations and create models through probing examples and counterexamples. They are expected to communicate their mathematical reasoning through the clear, concise, and correct use of mathematical symbolism and logical thinking.

Figure G-2
Draft Descriptions of the Achievement Levels
Prepared by the Original Level-setting Panel

Fourth-grade Draft Descriptions

BASIC. The Basic level signifies some evidence of conceptual and procedural understanding in the five NAEP content areas of Numbers and Operations; Measurement; Geometry; Data Analysis, Statistics, and Probability; and Algebra and Functions. Understanding simple facts and single-step operations are included at this level, as is the ability to perform simple computations with whole numbers. This level shows a partial mastery of estimation, basic fractions, and decimals relating to money or the number line; it shows an ability to solve simple real-world problems involving measurement, probability, statistics, and geometry. At this level, there is a partial mastery of tools such as four-function calculators and manipulatives (geometric shapes and rulers). Written responses are often minimal, perhaps with a partial response and lack of supportive information.

PROFICIENT. The Proficient level signifies consistent demonstration of the integration of procedural knowledge and conceptual understanding as applied to problem solving in the five NAEP content areas of Numbers and Operations; Measurement; Geometry; Data Analysis, Statistics, and Probability; and Algebra and Functions. The Proficient level indicates an ability to perform computation and estimation with whole numbers, to identify fractions, and to work with decimals involving money or the number line. Solving real-world problems involving measurement, probability, statistics, and geometry is an important part of this level. This level signifies the ability to use, as tools, four-function calculators, rulers, and manipulatives (geometric shapes). It includes the ability to identify and use pertinent/appropriate information in problem settings. The ability to make connections between and among skills and concepts emerges at this level. Clear and organized written presentations, with supportive information, are typical. And, there is an ability to explain how the solution was achieved.

ADVANCED. The Advanced level signifies the integration of procedural knowledge and conceptual understanding as applied to problem solving in the five NAEP content areas of Numbers and Operations; Measurement; Geometry; Data Analysis, Statistics, and Probability; and Algebra and Functions. This is evidenced by divergent and elaborate written responses. The Advanced level indicates an ability to solve multistep and nonroutine real-world problems involving measurement, probability, statistics, and geometry, and an ability to perform complex tasks involving multiple steps and variables. Tools are mastered, including four-function calculators, rulers, and manipulatives (geometric shapes). This level signifies the ability to apply facts and procedures by explaining *why* as well as *how*. Interpretations extend beyond obvious connections and thoughts are communicated clearly and concisely. At this level, logical conclusions can be drawn and complete justifications can be provided for answers and/or solution processes.

Eighth-grade Draft Descriptions

BASIC. Basic students should begin to describe objects, to process accurately and elaborate relationships, to compare and contrast, to find patterns, to reason from graphs, and to understand spatial reasoning. This level of partial mastery signifies an understanding of arithmetic operations on whole numbers, decimals, fractions, and percents, including estimation. Problems that are already set up are generally solved correctly, as are one-step problems. However, problems involving the use of available data, and determinations of what is necessary and sufficient to solve the problem, are generally quite difficult. Students should select appropriate problem-solving tools, including calculators, computers, and manipulatives (geometric shapes) to solve problems from the five content areas. Students should also be able to use elementary algebraic concepts

Figure G-2 (continued)
Draft Descriptions of the Achievement Levels
Prepared by the Original Level-setting Panel

and elementary geometric concepts to solve problems. This level indicates familiarity with the general characteristics of measurement. Students at this level may demonstrate limited ability to communicate mathematical ideas.

PROFICIENT. Proficient students apply mathematical concepts consistently to more complex problems. They should make conjectures, defend their ideas, and give supporting examples. They have developed the ability to relate the connections between fractions, percents, and decimals, as well as other mathematical topics. The Proficient level denotes a thorough understanding of the arithmetic operations listed at the Basic level. This understanding is sufficient to permit applications to problem solving in practical situations. Quantity and spatial relationships are familiar situations for problem solving and reasoning, and this level signifies an ability to convey the underlying reasoning skills beyond the level of arithmetic. Ability to compare and contrast mathematical ideas and generating examples is within the Proficient domain. Proficient students can make inferences from data and graphs; they understand the process of gathering and organizing data, calculating and evaluating within the domain of statistics and probability, and communicating the results. The Proficient level includes the ability to apply the properties of elementary geometry. Students at this level should accurately use the appropriate tools of technology.

ADVANCED. The Advanced level is characterized by the ability to go beyond recognition, identification, and application of mathematical rules in order to generalize and synthesize concepts and principles. Generalization often takes shape through probing examples and counterexamples and can be focused toward creating models. Mathematical concepts and relationships are frequently communicated with mathematical language, using symbolic representations where appropriate. Students at the Advanced level consider the reasonableness of an answer, with both number sense and geometric awareness. Their abstract thinking ability allows them to create unique problem-solving techniques and explain the reasoning processes they followed in reaching a conclusion. These students can probe through examples and counterexamples that allow generalization and description of assumptions with models and elegant mathematical language.

Twelfth-grade Draft Descriptions

BASIC. This level represents understanding of fundamental algebraic operations with real numbers, including the ability to solve two-step computational problems. It also signifies an understanding of elementary geometrical concepts such as area, perimeter, and volume, and the ability to make measurements of length, weight, capacity, and time. Also included at the Basic level is the ability to comprehend data in both tabular and graphical form and to translate between verbal, algebraic, and graphical forms of linear expression. Students at this level should be able to use a calculator appropriately.

PROFICIENT. This level represents mastery of fundamental algebraic operations and concepts with real numbers, and an understanding of complex numbers. It also represents understanding of polynomials and their graphs up to the second degree, including conic sections. The elements of plane, solid, and coordinate geometry should be understood at the Proficient level. The Proficient level includes the ability to apply concepts and formulas to problem solving. Students at this level should demonstrate critical thinking skills. The Proficient level also represents the ability to judge the reasonableness of answers and the ability to analyze and interpret

Figure G-2 (continued)
Draft Descriptions of the Achievement Levels
Prepared by the Original Level-setting Panel

data in both tabular and graphical form. Basic algebraic concepts, measurement, and constructive geometry concepts are mastered at this level.

ADVANCED. The Advanced level represents mastery of trigonometric, exponential, logarithmic, and composite functions, zeros and inverses of functions, polynomials of the third degree and higher, rational functions, and graphs of all of these. In addition, the Advanced level represents mastery of topics in discrete mathematics including matrices and determinants, sequences and series, and probability and statistics, as well as topics in analytic geometry. The Advanced level also signifies the ability to successfully apply these concepts to a variety of problem-solving situations.

Figure G-3
Revised Draft Descriptions of the Achievement Levels
Recommended by the Follow-Up Validation Panel

Revised Fourth-grade Draft Descriptions

BASIC. Basic students exhibit some evidence of conceptual and procedure understanding in the five NAEP content areas. At the fourth grade level, algebra and functions are treated in informal and exploratory ways often through the study of patterns. Basic students estimate and use basic facts to perform simple computations with whole numbers. These students show some understanding of fractions and decimals. They solve simple real world problems in all areas. These students use, although not always accurately, four-function calculators, rulers, and geometric shapes. Written responses are often minimal and lack supporting information.

PROFICIENT. Proficient students consistently integrate procedural knowledge and conceptual understanding as applied to problem solving in the five NAEP content areas. Using whole numbers they estimate, compute, and determine whether their results are reasonable. They have a conceptual understanding of fractions and decimals. Solving real world problems in all areas is important at this level. Proficient students appropriately use four-function calculators, rulers and geometric shapes. These students use problem solving strategies such as identifying and using appropriate information. [Problem-solving strategies include identification and use of appropriate information.] They present organized written solutions with supporting information and explain how they were achieved.

ADVANCED. Advanced students integrate procedural knowledge and conceptual understanding as applied to problem solving in the five NAEP content areas. They solve complex and non-routine real world problems in all areas. They have mastered the use of tools such as four-function calculators, rulers and geometric shapes. Advanced students draw logical conclusions and justify answers and solution processes by explaining the "why" as well as the "how." Interpretations extend beyond obvious connections and thoughts are communicated clearly and concisely.

Revised Eighth-grade Draft Descriptions

BASIC. Basic students exhibit evidence of conceptual and procedural understanding. These students compare and contrast, find patterns, reason from graphs, and understand spatial reasoning. This level of performance signifies an understanding of arithmetic operations, including estimation, on whole numbers, decimals, fractions, and percents. Students complete problems correctly with the help of structural prompts such as diagrams, charts, and graphs. As students approach the proficient level, they will solve problems involving the use of available data, and determine what is necessary and sufficient for a correct solution. Students use problem solving strategies and select appropriate tools, including calculators, computers, and manipulatives (geometric shapes) to solve problems from the five content areas. Students use fundamental algebraic and informal geometric concepts to solve problems. Students at this level demonstrate limited skills in communicating mathematically.

PROFICIENT. Proficient students apply mathematical concepts and procedures consistently to complex problems. They make conjectures, defend their ideas, and give supporting examples. They have developed the ability to relate the connections between fractions, percents, and decimals, as well as other mathematical topics, such as algebra and functions. The proficient level denotes a thorough understanding of the arithmetic

Figure G-3 (continued)
Revised Draft Descriptions of the Achievement Levels
Recommended by the Follow-Up Validation Panel

operations listed at the basic level. This understanding is sufficient to permit applications to problem solving in practical situations. Quantity and spatial relationships are familiar situations for problem solving and reasoning, and students at this level convey the underlying reasoning skills beyond the level of arithmetic. Proficient students compare and contrast mathematical ideas and generate their own examples. These students make inferences from data and graphs; they understand the process of gathering and organizing data, calculating, evaluating, and communicating the results within the domain of statistics and probability. Proficient students apply the properties of informal geometry, and accurately use the appropriate tools of technology.

ADVANCED. Advanced students go beyond recognition, identification, and application of mathematical rules in order to generalize and synthesize concepts and principles. Generalization often takes shape through probing examples and counter examples and can be used to create models. Mathematical concepts and relationships are frequently communicated with mathematical language, using symbolic representations where appropriate. Students at the advanced level consider the reasonableness of an answer, with both number sense and geometric awareness. Their abstract thinking allows them to create unique problem solving techniques and explain the reasoning processes they followed in reaching a conclusion. These students probe examples and counter examples that allow generalization and description of assumptions with models and elegant mathematical language.

Revised Twelfth-grade Draft Descriptions

BASIC. Basic students demonstrate procedural and conceptual knowledge in solving problems in the five NAEP content areas. They use estimation to verify solutions and determine the reasonableness of the results to real world problems. Algebraic and geometric reasoning strategies are used to solve problems. These students recognize relationships in verbal, algebraic, tabular, and graphical forms. Basic students demonstrate knowledge of geometric relationships as well as corresponding measurement skills. Statistical reasoning is applied to the organization and display of data and to reading tables and graphs. These students generalize from patterns and examples in the areas of algebra, geometry, and statistics. They communicate mathematical relationships and reasoning processes with correct mathematical language and symbolic representations. Calculators are used appropriately to solve problems.

PROFICIENT. Proficient students integrate mathematical concepts and procedures consistently to more complex problems in the five NAEP content areas. They demonstrate an understanding of algebraic reasoning, geometric and spatial reasoning, and statistical reasoning as applied to other areas of mathematics. They perform algebraic operations involving polynomials, justify geometric relationships, and judge and defend the reasonableness of answers in real world situations. These students analyze and interpret data in tabular and graphical form. Proficient students understand and use elements of the function concept in symbolic, graphical and tabular form. They make conjectures, defend their ideas, and give supporting examples.

ADVANCED. Advanced students consistently demonstrate the integration of procedural and conceptual knowledge, as well as the synthesis of ideas, in the five NAEP content areas. Advanced students understand the function concept, and they compare and apply the numeric. algebraic, and graphical properties of functions. They apply and connect their knowledge of algebra, geometry, and statistics to solve problems in more advanced areas of continuous and discrete mathematics. Advanced students formulate generalizations using examples and counter examples to create models. In communicating their mathematical reasoning, these students demonstrate clear, concise, and correct use of mathematical symbolism and logical thinking.

Figure G-4
Meeting Participants, NAEP Mathematics Achievement Level-setting
Original Meeting, St. Louis, Missouri, March 20-24, 1992

Marge Blizard
Blizard Professional Cleaning
Franklin, Connecticut

Christopher Chomyak
The Episcopal Church
Calais, Maine

Janet Green
Met Life
Crownsville, Maryland

Mary Norman
Dekalb County Board of Education
Decatur, Georgia

Janice Wamsley
Alcorn School System
Glen, Mississippi

Ronald Higgins
Walla Walla School District
Walla Walla, Washington

Leona Lee
Baltimore City Public Schools
Baltimore, Maryland

Lisa Bietau
USD 383 Manhattan Public Schools
Manhattan, Kansas

Marsha Davis
Alcorn County Schools
Corinth, Mississippi

Jean Bush Ragin
Patterson High School
Baltimore, Maryland

Bill Oldham
Harding University
Searcy, Arkansas

George Shell
Retired Principal
Draper, Utah

Marsha Stovey
Detroit Public Schools
Detroit, Michigan

Vance Morris
Dekalb County Board of Education
Atlanta, Georgia

Carol Ballentine
Duval County Schools
Jacksonville, Florida

Tami Harvey, ESD
Audiometric Technician
Burns, Oregon

Laurence Payne
Greater Houston Coalition
  for Educational Excellence
Houston, Texas

Cheryl Yunk
USD 383
Manhattan, Kansas

Kirby Gchachu
Zuni Public School District
Zuni, New Mexico

Corliss Hubert
Rutherford Board of Education
Englewood, New Jersey

Joyce Dunn
Alcorn County Schools
Corinth, Mississippi

Gloria Moran
Williams Junior High
Bridgewater, Massachusetts

Charles Jackson
Blairsville, Pennsylvania

Cassandra Turner
Internal Revenue Service
Miami, Florida

Figure G-4 (continued)
Meeting Participants, NAEP Mathematics Achievement Level-setting
Original Meeting, St. Louis, Missouri, March 20-24, 1992

Jack Deal
Bethel Park School District
Pittsburgh, Pennsylvania

Ninfa Rivera
Lyford CISD
Raymondville, Texas

Gerald Zeringue
Garrity Construction Company
Harvey, Louisiana

Linda Brown
Van Zile Elementary School
Detroit, Michigan

Judy Bibb
Lonoke High School
Cabot, Arkansas

David Rank
School District of Greenville
Greenville, South Carolina

John Sweeney
Freed-Haideman University
Henderson, Tennessee

Nancy Pejouhy
Woodstock Union High School
Woodstock, Vermont

Jim Trefzger
Parkland College
Champaign, Illinois

Joanne Greaver
Jefferson County Public Schools
Louisville, Kentucky

Ellie Cucinatto
Bridgewater Public Schools
Bridgewater, Massachusetts

Lillie Carr
Pender County Schools
Teachey, North Carolina

Eric Cain
IBM
Metairie, Louisiana

Phillip Stroup
Butler County MR/DD
Seven Mile, Ohio

Mike Gobel
Walla Walla School District
Walla Walla, Washington

Juanita Tietze
Retired Principal
Canton, Ohio

Bill Cramer, Jr.
Cramer & Mallon, Attorneys at Law
Burns, Oregon

Norma Newman
Yaleta Independent School District
El Paso, Texas

William Rickenbach
Bethel Park School District
Bethel Park, Pennsylvania

Violet Cosgrove
Retired
Glen Burnie, Maryland

Danny McDougal
Pre-Mc, Inc.
Allen, Oklahoma

Bill Anderson
Administration Eagle Union
Zionsville, Indiana

Dan Thompson
Thompson Construction Company
Trinidad, Colorado

Figure G-4 (continued)
Meeting Participants, NAEP Mathematics Achievement Level-setting
Original Meeting, St. Louis, Missouri, March 20-24, 1992

Nancy Gallagher
West Penn Power Company
Kittanning, Pennsylvania

William Hawes
The Hawes Company
Tucker, Georgia

Zhining Qin
Minnesota Department of Education
St. Paul, Minnesota

Charles McGee
Greenville County School District
Greenville, South Carolina

Barbara Bayne
Greenville County School District
Greenville, South Carolina

Landa McLaurin
Baltimore City Schools
Baltimore, Maryland

Nancy Potempa
St. Xavier University
Mokena, Illinois

Florencetine Jasmin
Baltimore City Public Schools
Baltimore, Maryland

Jeane Joyner
Dept. of Public Instruction
Raleigh, North Carolina

Carolyn Craig
State Department of Education
Jackson, Mississippi

Mary Bennion
SFUSD
San Francisco, California

Bill Eyestone
Walla Walla School District
Walla Walla, Washington

Mary Gahn
New York Board of Education
Westbury, New York

Barbara Faltz-Jackson
Baltimore City Public Schools
Baltimore, Maryland

Florence Kelly
Manville Board of Education
Manville, New Jersey

Philip Brach
Univ. of the District of Columbia
Washington, D.C.

Larry Brown
Oil industry (Self-Employed)
Allen, Oklahoma

W. Garry Quast
Slippery Rock University
Slippery Rock, Pennsylvania

Carl Springfels
Consultant (Self-Employed)
Miami Shores, Florida

Anna Maria Golan
Santa Ana Unified
Fountain Valley, California

Ricardo Suarez
Lyford CISD
Raymondville, Texas

Figure G-5
Meeting Participants, NAEP Mathematics Achievement Level-setting Follow-up Validation Meeting, Nantucket, Massachusetts, July 17-19, 1992

Charles Allen
Michigan Department of Education
Lansing, Michigan

Linda Brown
Van Zile Elementary School
Clinton Township, Michigan

Ellie Cucinatto
Bridgewater Public Schools
Bridgewater, Massachusetts

Jack Deal
Bethel Park School District
Pittsburgh, Pennsylvania

Paula Duckett
River Terrace Community School Board
Washington, DC

Edward Esty
SRI International
Washington, D.C.

Barbara Faltz-Jackson
Baltimore Public Schools
Baltimore, Maryland

Joan Ferini-Mundy
University of New Hampshire
Durham, New Hampshire

Marilyn Hala
National Council of Teachers of Mathematics
Washington, D.C.

Florence Kelly
Largo Public Schools
Largo, Florida

Henry Kepner
University of Wisconsin at Milwaukee
Milwaukee, Wisconsin

Charles McGee
Greenville Public Schools
Greenville, South Carolina

Landa McLaurin
Baltimore City Schools
Baltimore, Maryland

Gloria Moran
Williams Junior High School
Bridgewater, Massachusetts

Jo Ann Mosier
Kentucky Department of Education
Frankfort, Kentucky

Mary Norman
DeKalb County Board of Education
Decatur, Georgia

David Rank
Greenville Public Schools
Greenville, South Carolina

Sharon Steglein
Minnesota Department of Education
St. Paul, Minnesota

APPENDIX H

REANALYSIS OF THE 1990 TRIAL STATE ASSESSMENT DATA

Appendix H

Reanalysis of the 1990 Trial State Assessment Data

John Mazzeo
Educational Testing Service

As has been the case since 1984, the estimation of proficiency scale distributions for the 1992 national and Trial State Assessments was carried out using the plausible values methodology described in Mislevy (1991) and in Chapter 8 of the current report. The methodology is implemented using Sheehan's (1985) MGROUP computer program and involves the estimation of a multivariate linear model for the regression of proficiency (θ) on a large number of predictor variables related to examinee background characteristics and instructional experience. The version of the program used in 1990, based on the EM algorithm described in Mislevy (1985), used Monte Carlo integration procedures to estimate the parameters of linear regression model. Subsequent to the 1990 assessment, these estimation procedures were improved by the introduction of analytic integration procedures and the incorporation of higher-order asymptotic corrections to estimates of examinee means and posterior variances (Thomas, 1992).

Preliminary research with simulated data and experience with selected reanalyses of previously reported 1990 NAEP data sets (both national and Trial State assessments) suggested that results from the revised program would differ from those obtained with earlier versions of MGROUP to a degree that was not ignorable. The 1990 estimates of the correlations between scales had been substantially attenuated. For example, estimates of correlations among the mathematics subscales (conditional on the full set of background variables used for the analysis) obtained with the version of MGROUP used in 1990 were typically in the .15 to .25 range across all states that participated in that year's Trial State Assessment. When the 1990 data were reanalyzed with the revised MGROUP procedures used for the 1992 Trial State Assessment, these same correlations were typically found to be in the .85 to .95 range.

The underestimation of subscale correlations had little impact on the accuracy of results (means, standard deviations, and percentiles) reported for the five NAEP content area scales. In addition, this underestimation had little effect on the means reported for the mathematics composite scale. However, the composite scale is a weighted average of the results from each of its constituent scales and, as such, the standard deviation of the composite is partly a function of the interscale correlations. Consequently, composite scale variability had been underestimated for both the national and Trial State Assessments and this attenuation resulted in underestimates of the percentages of examinees outside the more extreme NAEP anchor points.

Plans for the 1992 mathematics assessment called for the use of the revised estimation procedures. However, the use of such procedures for the 1992 analyses alone would make accurate comparisons to 1990 difficult, if not impossible, for certain composite scale statistics. In order to maintain the integrity of the 1990 NAEP mathematics scales for trend analysis, a decision was made to reanalyze the 1990 results for both the national and Trial State assessments and to report the revised 1990 figures in conjunction with the 1992 results. The reanalyses involved only those aspects related to the MGROUP procedure. The item parameter estimates from the 1990 national and Trial State assessments were *not* re-estimated. However, the estimation of conditioning models and the generation of plausible values were redone using the 1990 data and applying the same version of the MGROUP program used for the 1992 assessment. In all other respects, the 1990 reanalyses involved procedures nearly identical[1] to those that produced the original 1990 results (Yamamoto & Jenkins, 1992, section 13.2.6; Mazzeo, 1992, section 10.5).

Resetting the Origin and Unit of the 1990 Reporting Scales

The reanalysis of the 1990 results engendered slightly different scaling transformation constants—that is, the constants that linearly transform NAEP results from the metric in which they are estimated (θ_{est} with approximate mean 0 and standard deviation of 1) to the proficiency metric in which they are reported (θ_{rep} with, in most cases, mean of 250.5 and standard deviation of 50. The procedures used to obtain scaling transformation constants for the revised 1990 results are described below.

Scaling Transformations for the 1990 National Mathematics Assessment

The revised 1990 scaling transformations were obtained by applying the same procedures originally used in 1990 (Yamamoto & Jenkins, 1992) to the revised 1990 results.

For the Numbers and Operations, Measurement, Geometry, and Algebra and Functions scales the procedure was as follows:

1) Separate estimates of the mean and variance of each scale were obtained in the national θ-metric for the winter half-sample of each of the three age/grade cohorts. These estimates were obtained using final sampling weights. The reason for using only the winter half-sample was to center the scale and establish its unit in terms of the most sensible reference population against which to compare the Trial State Assessment. Participants in the Trial State Assessment were tested during the same time period as the winter half-sample of the 1990 national assessment (January through March of 1990). Note, however, that this national standardization population is still not directly comparable to the Trial State Assessment population in that a) it contains age-eligible as well as grade-eligible examinees, b) contains private-school as well as public-school examinees, and c) contains examinees from states that did not participate in Trial State Assessment. It should also be noted that the national sample to which the Trial State Assessment is compared in the 1990 composite and state reports is not identical

---

[1] A slight change to the program used to obtain principal components resulted in one additional component being included as a predictor variable in the conditioning model for each age/grade cohort of the national assessment and for each jurisdiction of the Trial State Assessment.

to the national standardization sample. The national reporting comparison sample excluded age-eligible and private-school examinees.

2) The estimates from step 1 were combined to produce an estimate of the overall mean and standard deviation for a population of students tested in the winter and consisting of equal numbers of students from each age/grade cohort. The estimate of the overall mean was simply the unweighted average of the means for the three cohorts. The estimate of the standard deviation was obtained as the square root of the sum of the unweighted average of the within-cohort variances and the variance of the between-cohort means.

3) Constants were then derived that linearly transform the overall mean and standard deviation obtained in step 2 to 250.5 and 50, respectively. These values, which are consistent with previous NAEP practice, represent the mean and standard deviation of estimated true scores for the combined population of all three cohorts on a hypothetical test. The test consists of 500 items, equally-spaced between -4.99 and 4.99 on the θ_m scale, which follow a 1-parameter logistic model with a discrimination parameter of 1.5 (see Beaton, 1987, page 384).

Sufficient items to produce a Data Analysis, Statistics, and Probability scale were present for only two of the three age/grade cohorts. Therefore, a slightly different scale transformation procedure was used for this scale in 1990 and with the revised 1990 results. The procedure was as follows:

1) For the Data Analysis scale, means were obtained in the national θ metric for the winter half-samples of the grade 8/age 13 and grade 12/age 17 cohorts.

2) For the other four scales discussed above, means were obtained in the reporting metric for these same cohort.

3) For each cohort, a weighted average of the means obtained in step 2 was produced. The weights used were those employed in forming the 1990 composite scale at the corresponding grade, renormalized to sum to 1.

4) Constants were then derived that mapped the θ metric means obtained in step 1 to the weighted averages produced in step 3.

The procedure used to reset the Estimation scale was identical to that used to set the metric for the Numbers and Operations, Measurement, Geometry, and Algebra and Functions scales, the only difference being the use of both the winter and spring age/grade samples. The decision to use both half-samples was based on two considerations. First, because the estimation items had not been administered in the 1990 Trial State Assessment there was less need to center the scales in terms of a comparable reference sample. Second, the estimation items were administered to a separate, and smaller, national sample than was the BIB portion of the main assessment (which contains the items for the other scales). Restricting data to the winter-half sample for the estimation scale would have resulted in smaller than desired sample sizes for the standardization.

The original 1990 transformations and their corresponding revised values are given in Table H-1. The transformations are of the form $\theta_{rpt} = k_1 + k_2(\theta_{est})$.

Scaling Transformations for the 1990 Trial State Assessment in Mathematics

The scaling transformations for the 1990 Trial State Assessment were also redone using identical procedures to those reported in Mazzeo (1991), but applied to the reestimated 1990 results. The procedure, which was intended to equate the metrics of the Trial State Assessment scales to their corresponding national scales, was as follows:

1) Means and standard deviations in the Trial State Assessment θ-metric were obtained for each of the five scales for the aggregate sample of Trial State Assessment examinees from all participating jurisdictions with the exception of Guam and the Virgin Islands. These latter two participants were excluded because no data from corresponding PSUs were available from the national assessment. Final sampling weights provided by Westat were used in producing the necessary sample moments.

2) Corresponding means and standard deviations for these scales in the national θ-metric were also obtained for a restricted sample of the national assessment, referred to as the State Aggregate Comparison (SAC) sample. The SAC sample consisted of public-school, grade-eligible students from only those states that participated in the 1990 Trial State Assessment. Special weights were provided by Westat to enhance comparability of this sample to the aggregate of the Trial State Assessment on which the means and standard deviations from step 1 were obtained.

3) A set of transformation constants were obtained linking the two θ-metrics by setting means and standard deviations equal.

4) These constants were then concatenated with the appropriate constants in Table 1 to produce a final set of revised Trial State Assessment scaling constants.

The original and revised Trial State Assessment scaling constants are given in Table H-2.

Comparisons of Original and Revised Results for the 1990 Trial State Assessment

In the vast majority of cases, differences between the 1990 results originally reported and the revised results are extremely small. For example, the revised state means and means for subgroups on each of the five content area scales and the composite scales are, for the most part, within 1 standard error of the originally reported values. Figure H-1 provides plots of the revised state means versus original means for the Numbers and Operations scale (the scale for which examinees were administered the greatest number of items), the Data Analysis, Statistics, and Probability scale (the scale for which examinees were administered the fewest items), and

Table H-1
Original and Revised Scaling Transformation Constants for the 1990 National Mathematics Scales

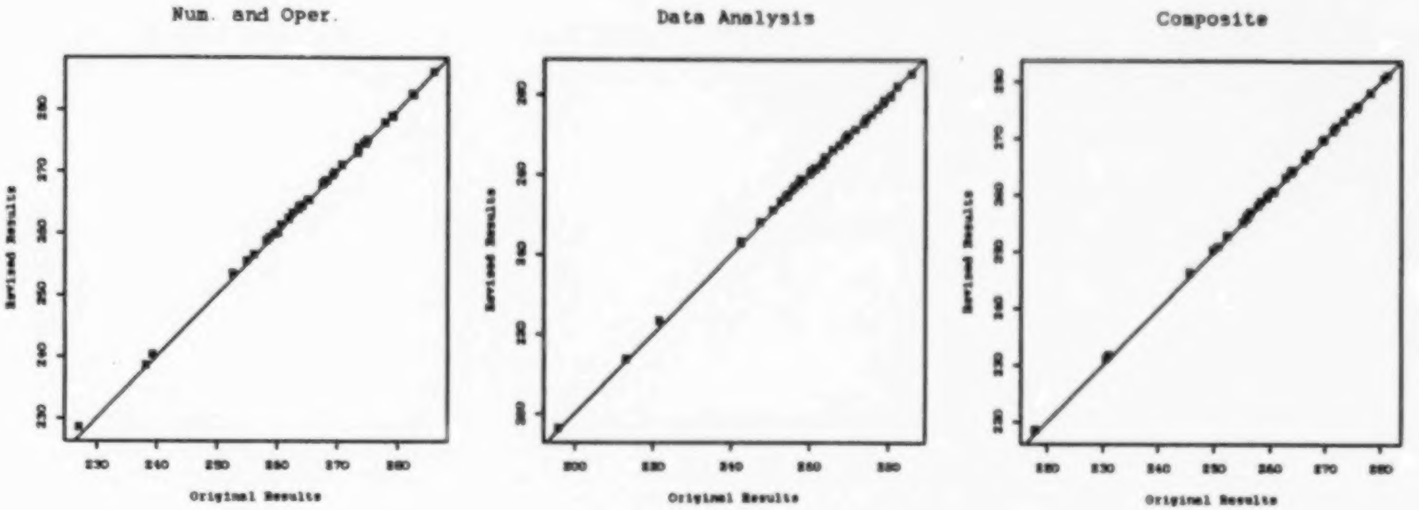
Scale	Original Transformation		Revised Transformation	
	k_1	k_2	k_1	k_2
Numbers & Operations	251.72	50.35	251.82	50.12
Measurement	252.59	49.99	252.18	49.79
Geometry	252.58	49.15	252.79	48.81
Data Analysis	274.08	44.47	274.19	43.68
Algebra & Functions	252.01	50.34	251.90	49.99
Estimation	249.52	49.62	249.51	49.53

Table H-2
Original and Revised Scaling Transformation Constants for the 1990 Trial State Assessment Mathematics Scales

Scale	Original Transformation		Revised Transformation	
	k_1	k_2	k_1	k_2
Numbers & Operations	265.28	36.14	265.87	35.35
Measurement	256.69	43.92	256.77	42.18
Geometry	258.90	35.81	259.23	34.79
Data Analysis	259.64	44.84	260.21	43.09
Algebra & Functions	259.71	38.12	260.27	37.43

Figure H-1

Revised 1990 State Means Plotted Against Originally Reported 1990 State Means for Three Scales



for the composite scale for each of the 40 participants. As is evident from the figure, differences were quite small and the rank ordering of the states was unaffected. Essentially identical results were obtained for the scales not shown.

As discussed earlier, the principal difference between the results produced by the MGROUP program used in 1990 and that used in 1992 involved estimates of the within-jurisdiction standard deviations and the proportions of students exceeding NAEP anchor points. Figure H-2 contains a stem-and-leaf display of the ratio of the revised 1990 composite scale standard deviation estimates to the original estimates for each of the 40 participants. The results originally reported were consistently lower than the revised results, with the underestimation ranging from 6 to 16 percent.

Figure H-2
Stem-and-Leaf Display* of
Ratios of Composite Standard Deviations (Revised 1990/Original 1990)

N = 40, Median = 1.1075, Quartiles = 1.0945, 1.121
Decimal point is 2 places to the left of the colon

```
 1    1  106 : 3
 2    1  107 : 1
 8    6  108 : 224558
13    5  109 : 13689
     10  110 : 2356777899
17    5  111 : 01589
12    8  112 : 00235578
 4    2  113 : 02
 2    1  114 : 1
 1    0  115 :
 1    1  116 : 3
```

---

* The first column of numbers shows observation depths; the second column shows the number of observations; the remainder of the figure contains the stem-and-leaf display.

Table H-3 provides differences (original 1990 results minus revised 1990 results) in the estimated proportions at or above each NAEP anchor point for each of the participating jurisdictions. Differences were typically on the order of 1 to 2 percent and none exceeded 4 percent. In general, the percentages at or above the higher anchor points were slightly underestimated, while the percentages at or above the lowest anchor point were slightly overestimated. These differences were a direct result of the underestimation of the correlations between scales and the resulting underestimation in the composite scale standard deviations.

Table H-3
Differences (Original 1990 Results Minus Revised 1990 Results) in Estimated Percentages At or Above Each NAEP Anchor Point for Each Participating Jurisdiction

Jurisdiction	NAEP Anchor Levels			
	200	250	300	350
Alabama	1.8	-1.1	-1.8	-0.2
Arizona	2.1	-0.3	-2.2	-0.3
Arkansas	1.8	-0.6	-2.0	-0.1
California	2.4	-1.0	-1.4	-0.3
Colorado	1.6	1.5	-2.2	-0.3
Connecticut	1.3	1.0	-2.3	-0.4
Delaware	1.9	-0.7	-0.8	-0.2
District of Columbia	3.1	-3.1	-0.6	-0.1
Florida	2.5	-1.5	-1.3	-0.2
Georgia	1.5	-0.9	-1.5	-0.4
Guam	2.6	-3.8	-0.7	-0.1
Hawaii	2.8	-2.1	-1.1	-0.3
Idaho	0.7	2.7	-2.6	-0.2
Illinois	1.7	0.9	-2.5	-0.4
Indiana	1.0	1.2	-2.3	-0.4
Iowa	0.5	3.2	-3.1	-0.4
Kentucky	1.9	-0.6	-2.1	-0.2
Louisiana	2.2	-2.6	-1.3	-0.2
Maryland	1.7	-0.6	-2.0	-0.4
Michigan	1.1	0.3	-2.2	-0.6
Minnesota	0.7	2.7	-2.1	-0.8
Montana	0.3	2.6	-2.4	-0.6
Nebraska	1.0	2.0	-2.6	-0.5
New Hampshire	1.0	1.8	-2.3	-0.3
New Jersey	1.1	0.9	-1.7	-0.5
New Mexico	1.9	-1.1	-1.9	-0.2
New York	2.0	-0.4	-1.9	-0.6
North Carolina	2.4	-1.9	-1.4	-0.1
North Dakota	0.3	2.4	-2.2	-0.8
Ohio	1.2	0.4	-2.0	-0.3
Oklahoma	1.4	0.4	-3.0	-0.1
Oregon	1.5	1.6	-1.6	-0.5
Pennsylvania	1.2	-0.2	-1.8	-0.4
Rhode Island	0.9	-0.5	-1.9	-0.2
Texas	1.8	-0.5	-2.0	-0.4
Virginia	1.1	-0.4	-1.5	-0.6
Virgin Islands	1.9	-3.0	-0.2	-0.0
West Virginia	2.2	-0.9	-1.4	-0.2
Wisconsin	0.6	2.2	-2.7	-0.3
Wyoming	0.8	2.2	-3.1	-0.1

REFERENCES CITED IN TEXT

REFERENCES CITED IN TEXT

Abt Associates. (1991). *Prospects: The National Longitudinal Study of Chapter I children* (Final progress report for design contract No. LC89027001). Chicago, IL: Author.

Andersen, E. B. (1980). Comparing latent distributions. *Psychometrika, 45,* 121-134.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508 - 600). Washington, DC: American Council on Education.

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics, 17,* 191-204.

Beaton, A. E., & Johnson, E. G. (1992). Overview of the scaling methodology used in the National Assessment. *Journal of Educational Measurement, 26(2),* 163-175.

Beaton, A. E., & Johnson, E. G. (1990). The average response method of scaling. *Journal of Educational Statistics, 15,* 9-38.

Bourque, M. L., & Garrison, H. H. (1991). *The levels of mathematics achievement. Vol. I, national and state summaries.* Washington, DC: National Assessment Governing Board.

Burke, J., Braden, J., Hansen, M., Lago, J., Tepping, B. (1987). *National Assessment of Educational Progress – 17th year sampling and weighting procedures. Final report.* Rockvill, MD: Westat, Inc.

Cochran, W. G. (1977). *Sampling techniques.* New York, NY: John Wiley & Sons.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York, NY: Academic Press.

Curry, L. (1987, April). *Group decision process in setting cut-off scores.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, 39,* 1-38.

Educational Testing Service (1987). *ETS standards for quality and fairness.* Princeton, NJ: Author.

Engelen, R. J. H. (1987). *Semiparametric estimation in the Rasch model.* Research Report 87-1. Twente, the Netherlands: Department of Education, University of Twente.

Fitzpatrick, A. R. (1989). Social influences in standard-setting: The effects of social interaction on group judgments. *Review of Educational Research, 59,* 315-328.

Friedman, C. B., & Ho, K. T. (1990, April). *Interjudge consensus and intrajudge consistency: Is it possible to have both on standard setting?* Paper presented at the annual meeting of the National Council for Measurement in Education, Boston, MA.

Hambleton, R. K., & Bourque, M. L. (1991). *The levels of mathematics achievement. Vol. II, technical report.* Washington, DC: National Assessment Governing Board.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). *Understanding robust and exploratory data analysis.* New York, NY: John Wiley & Sons.

Hoijtink, H. (1991). *Estimating the parameters of linear models with a latent dependent variable by nonparametric maximum likelihood.* Research Bulletin HB-91-1040-EX. Groningen, The Netherlands: Psychological Institute, University of Groningen.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity.* Hillsdale, NJ: Erlbaum.

Johnson, E. G., & Allen, N. L. (1992). *The NAEP 1990 technical report* (No. 21-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Johnson, E. G., & Rust, K. F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics, 17,* 175-190.

Johnson, E. G., & Zwick, R. (1990). *Focusing the new design: The NAEP 1988 technical report* (No. 19-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Keyfitz, N. (1951). Sampling with probability proportional to size; adjustment for changes in probabilities. *Journal of the American Statistical Association, 46,* 105-109.

Kish, L. (1965). *Survey sampling.* New York: John Wiley & Sons.

Laird, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association, 73,* 805-811.

Lindsey, B., Clogg, C. C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association, 86,* 96-107.

Little, R. J. A., & Rubin, D. B. (1983). On jointly estimating parameters and missing data. *American Statistician, 37,* 218-220.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York, NY: John Wiley & Sons.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mazzeo, J. (1991). Data analysis and scaling. In S. L. Koffler, *The technical report of NAEP's 1990 Trial State Assessment program* (No. ST-21-01). Washington, DC: National Center for Education Statistics.

Mazzeo, J., Johnson, E. G., Bowker, D., & Fong, Y. F. (1992). *The use of collateral information in proficiency estimation for the Trial State Assessment*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177-196.

Mislevy, R. J. (1990). Scaling procedures. In E.G. Johnson and R. Zwick, *Focusing the new design: The NAEP 1988 technical repor* (No. 19-TR-20). Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80*, 993-997.

Mislevy, R. J., & Bock, R. D. (1982). *BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville, IN: Scientific Software.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17(2)*, 131-154.

Mislevy, R. J., & Sheehan, K. M. (1987). Marginal estimation procedures. In A.E. Beaton, *Implementing the new design: The NAEP 1983-84 technical report* (No. 15-TR-20). Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

Mislevy, R. J., & Stocking, M. L. (1987). *A consumer's guide to LOGIST and BILOG*. (ETS Research Report 87-43). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Wu, P-K. (1988). *Inferring examinee ability when some item responses are missing* (ETS Research Report RR-88-48-ONR). Princeton, NJ: Educational Testing Service.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16(2)*, 159-176.

Muraki, E., & Bock, R. D. (1991). *PARSCALE: Parameter scaling of rating data*. Chicago, IL: Scientific Software, Inc.

National Assessment Governing Board (1989). *Setting achievement goals on NAEP, a draft policy statement.* Washington, DC: Author.

National Assessment of Educational Progress (1992). *1992 policy information framework.* Princeton, NJ: Educational Testing Service.

National Assessment of Educational Progress (1988). *Mathematics objectives, 1990 assessment.* Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

National Assessment of Educational Progress (1987). *Mathematics objectives: 1985-86 assessment.* Princeton, NJ: Educational Testing Service.

National Council of Teachers of Mathematics (1987). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

Petersen, N. (1988). *DIF procedures for use in statistical analysis.* Internal memorandum.

Potter, F. (1988). Survey of procedures used to control extreme sampling weights. *Proceedings of the Section on Survey Research Methods* (pp. 453-458). Washington, DC: American Statistical Association.

Rogers, A. M. (1991). *NAEP-MGROUP: Enhanced version of Sheehan's software for the estimation of group effects in multivariate models* [Computer program]. Princeton, NJ: Educational Testing Service.

Rubin, D. B. (1991). EM and beyond. *Psychometrika, 56,* 241-254.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons.

Rust, K. R., Bethel, J., Burke, J., & Hansen, M. (1990). *1988 National Assessment of Educational Progress sampling and weighting procedures. Final report.* Rockville, MD: Westat, Inc.

Rust, K. R., & Bryant, E. (1991). *The IEA Reading and Literacy Study design and implementation: National and international perspectives, population definitions and sample design.* Presented at the annual meeting of the American Educational Research Association, Chicago, Illinois.

Rust, K. R., Burke, J., Fahimi, M., & Wallace, L. (1992). *1990 National Assessment of Educational Progress sampling and weighting procedures. Part 2 - National Assessment.* Rockville, MD: Westat, Inc.

Sheehan, K. M. (1985). *M-GROUP: Estimation of group effects in multivariate models* [Computer program] Princeton, NJ: Educational Testing Service.

Smith, R. L., & Smith, J. K. (1988). Differential use of item information by judges using Angoff and Nedelsky procedures. *Journal of Educational Measurement, 25*, 259-274.

Stokes, L. (1990). A comparison of truncation and shrinking of sample weights. *Proceedings of the Annual Research Conference* (pp. 463-471). Washington, DC: U.S. Bureau of the Census.

Tanner, M., & Wong, W. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association, 82*, 528-550.

Thomas, N. (1992). *Higher order asymptotic corrections applied in an EM algorithm for estimating educational proficiencies*. Unpublished manuscript.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley Publishing Co.

Wainer, H. (1974). The suspended rootogram and other visual displays: An empirical validation. *The American Statistician, 28(4)*, 143-145.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-202.

Wingersky, M., Kaplan, B. A., & Beaton, A. E. (1987). Joint estimation procedures. In A. E. Beaton, *Implementing the new design: The NAEP 1983-84 technical report*. (No 15-TR-20) Princeton, NJ: National Association of Educational Progress, Educational Testing Service.

Yamamoto, K., & Jenkins, F. (1992). Data analysis for the mathematics assessment. In E. G. Johnson & N. L. Allen, *The NAEP 1990 technical report* (No. 21-TR-20). Washington, DC: National Center for Education Statistics.

Yamamoto, K., & Mazzeo, J. (1992). Item response theory scale linking in NAEP. *Journal of Educational Statistics, 17(2)*, 155-173.

Zieky, M. (1993). Practical questions in the use of DIF statistics. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Zwinderman, A. H. (1991). Logistic regression Rasch models. *Psychometrika, 56*, 589-600.